InstructAV: Instruction Fine-tuning Large Language Models for Authorship Verification

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable proficiency in a wide range of NLP tasks. However, when it comes to authorship verification (AV) tasks, which involve determining whether two given texts share the same authorship, even advanced models like ChatGPT exhibit notable limitations. This paper introduces a novel approach, termed InstructAV, for authorship verification. This approach utilizes LLMs in conjunction with a parameter-efficient fine-tuning (PEFT) method to simultaneously improve accuracy and explainability. The distinctiveness of InstructAV lies in its ability to align classification decisions with transparent and understandable explanations, representing a significant progression in the field of authorship verification. Through comprehensive experiments conducted across various datasets, InstructAV demonstrates its state-of-the-art performance on the AV task, offering high classification accuracy coupled with enhanced explanation reliability.

1 Introduction

Authorship verification (AV) aims to determine whether two texts are written by the same author (Halvani et al., 2019). This task has far-reaching applications in fields such as forensics, literature, and digital security, where accurate identification of the author is crucial (Stamatatos, 2016).

Historically, the domain of AV has predominantly utilized traditional models focusing on stylistic and linguistic features (Seroussi et al., 2011; Bevendorff et al., 2019), alongside machine learning-based techniques (Zhang et al., 2018; Saedi and Dras, 2021). Traditional stylometric approaches harness characteristics such as word lengths, sentence lengths, and function word usage to differentiate between the distinct writing styles of various authors. In contrast, machine learning methods, especially those employing deep learning architectures like BERT (Devlin et al., 2018) and RoBERTa (Jones et al., 2022), have been employed to discern complex patterns in writing styles. These techniques often exceed the efficacy of stylometric methods in discerning authorship, as evidenced by recent advancements (Konstantinou et al., 2022; Valdez-Valenzuela et al., 2023). This shift towards sophisticated machine learning methodologies marks a significant evolution in the field of AV (Zheng and Jin, 2023).

Despite their advancements, existing AV models primarily focus on binary classification, often lacking explanatory insights. The importance of explainability cannot be overstated, as it enables a deeper understanding of the model's decision-making processes and bolsters confidence in the results produced. Moreover, the absence of explainability in these models presents significant challenges in identifying and addressing potential biases inherent within their framework. In light of these limitations, AV models must not only achieve high accuracy, but also embody a level of transparency and interpretability. This dual requirement underscores the need for models that are both effective in their predictive capabilities and accessible in their decision-making processes.

This paper introduces the InstructAV framework, a novel approach for AV tasks, designed to not only accurately determine if two texts share the same authorship but also to provide robust linguistic explanations for the AV outcomes. The InstructAV model is distinguished by its integration of explainability into the classification process, establishing a link between its predictive outcomes and the accompanying explanatory information. Our extensive experimentation on three distinct AV datasets highlights the efficacy of the InstructAV framework. This framework excels in achieving superior accuracy in AV tasks and further distinguishes itself by generating coherent and credible explanations that support its predictions.

Our contributions can be summarized as follows:

(i) We propose the InstructAV framework for AV tasks to accurately determine whether two texts share the same author and to furnish robust linguistic explanations for the AV outcomes. (ii) We have curated three instruction-tuning datasets, each accompanied by dependable linguistic explanations for AV tasks. These datasets are intended to serve as valuable resources for advancing research in this field[1]. (iii) Both automated and human evaluation results demonstrate the effectiveness of the InstructAV in providing precise AV predictions and reliable linguistic explanations.

2 Related Work

2.1 Authorship Verification

In the last two decades, AV has evolved significantly, transitioning from traditional methods focusing on linguistic features like spelling and style to machine learning techniques (Boenninghoff et al., 2019). However, traditional machine learning, such as support vector machines, showed limited effectiveness (Konstantinou et al., 2022). Recent advancements involve contextual embeddings from language models like BERT, T5, and MPNET (Devlin et al., 2018; Raffel et al., 2020; Song et al., 2020), and further studies have explored graph convolutional networks and BiLSTM with attention mechanisms (Valdez-Valenzuela et al., 2023; Sun et al., 2023).

Nevertheless, the earlier neural network-based methods like Bert and MPNET are primarily designed for classification tasks, and inherently lack the capability to provide explanations for their decisions. With the proliferation of large language models (LLMs), Hung et al. (2023) introduced PromptAV, an innovative technique that capitalizes on LLMs for AV. This method employs step-by-step stylometric explanation prompts to bolster interpretability. Their research reveals that PromptAV surpasses both chain of thought (CoT) prompting and PS+ prompting baselines in performance, marking a significant stride in the utilization of LLMs for AV tasks.

While PromptAV was able to generate explanations for the AV predictions, the model is largely limited to a few-shot demonstration, which limits the consistency and relevance of the generated explanations. To overcome this gap, we propose InstructAV, which utilizes a instruction fine-tuning approach to improve both the classification and explanation of the AV task.

2.2 Parameter-Efficient Fine-Tuning of Large Language Models

The rise of LLMs like GPT-3 has catalyzed the development of innovative fine-tuning methodologies designed to overcome the constraints of computational resources and memory usage. Central to this development is Parameter-Efficient Fine-Tuning (PEFT), which focuses on selectively adjusting a subset of a model's parameters to meet the specific demands of various downstream applications (Ding et al., 2023; Liu et al., 2022). Within the realm of PEFT, adapters, notably compact and trainable modules integrated into pre-trained networks, have gained prominence. These adapters enable targeted model specialization for specific tasks or datasets without necessitating comprehensive network retraining (Hu et al., 2023).

The Low-Rank Adaptation (LoRA) adapter is a key example of PEFT, fine-tuning model weight matrices with low-rank modifications to enhance task-specific trainability while preserving model strengths (Hu et al., 2021). Another method, Prefix-Tuning, adds task-specific parameters to transformer layers for task alignment (Li and Liang, 2021). These PEFT methods combine adaptability with computational efficiency, allowing advanced LLMs to be used in resource-limited settings. LoRA, in particular, maintains LLMs' general capabilities while enabling task-specific fine-tuning, marking a significant field advancement (Ding et al., 2023). In our study, we fine-tune with LoRA outperform to perform the AV tasks.

3 Methodology

The overview of the InstructAV framework, as depicted in Figure 1, comprises three primary steps: data collection, consistency verification, and the fine-tuning of LLMs using the LoRA method (Hu et al., 2021). Initially, the framework focuses on the aggregation of explanatory data for AV samples. This process capitalizes on the inherent presence of binary classification labels within existing AV datasets. Following this, a stringent quality control protocol is implemented, aimed at verifying the alignment and consistency of the explanations with the corresponding classification labels. The final stage involves the synthesis of instruction-tuning data, which is a fusion of the gathered classifica-

[1]The code and datasets will be released in camera-ready.

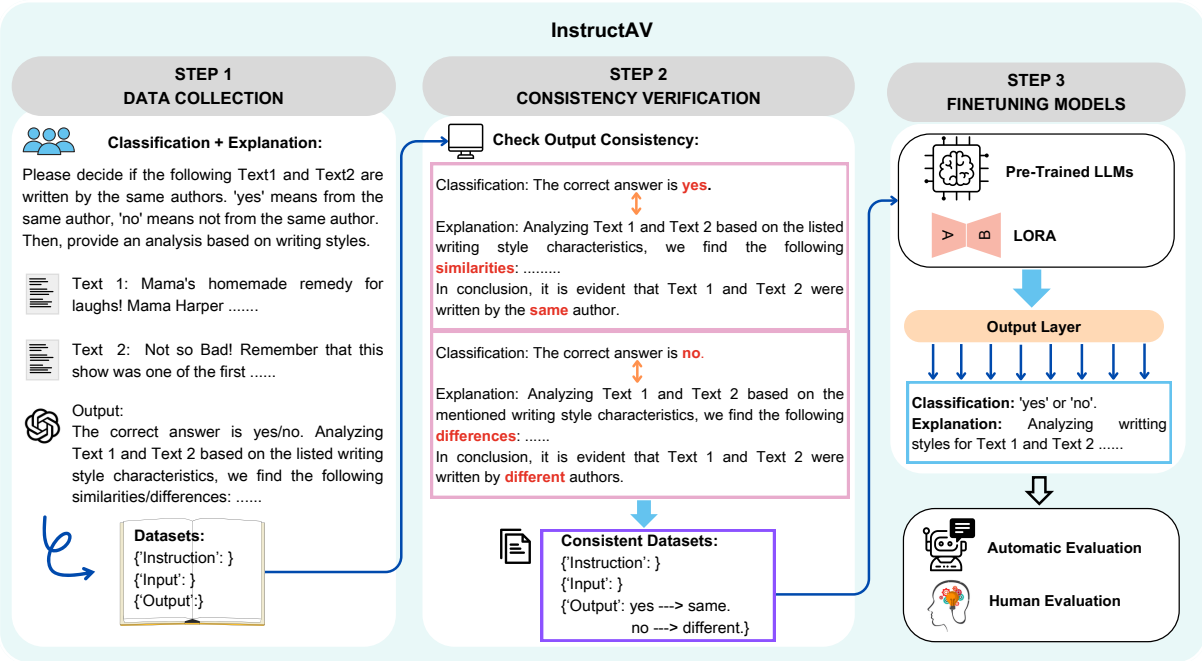

Figure 1: An illustration of InstructAV's architectures.

tion labels and their associated explanations. This composite data then serves as the foundation for fine-tuning LLMs in conjunction with the LoRA adaptation technique. This approach ensures that the LLMs are not only accurately fine-tuned for the AV task but also enhanced in their capacity to provide coherent and reliable explanations for their predictions. The subsequent sections present the details of each component in InstructAV.

3.1 Explanation Data Collection

To augment the explanatory capabilities of the InstructAV model, particularly in generating dependable explanations for AV predictions, we initiated a comprehensive collection of explanations using ChatGPT. This process involved three datasets widely used in AV studies: the IMDB dataset (Seroussi et al., 2014), the Twitter dataset (Schwartz et al., 2013), and the Yelp Reviews dataset[2]. This selection strategically covers various dimensions of textual data, thereby ensuring a diverse and comprehensive analysis.

The IMDB dataset, characterized by its longer text length (averaging 303 words), exemplifies long-form content. In contrast, the Twitter dataset, with an average text length of 16 words, epitomizes short-form content. The Yelp Reviews dataset, averaging 154 words per text, represents medium-length content. These datasets, initially curated for authorship attribution tasks, were adapted for our study. We extracted 11,000 samples from each dataset to facilitate a robust AV evaluation. Each sample comprises two texts, possibly authored by the same individual or different authors.

For generating explanations, we employed ChatGPT with few-shot prompts, focusing on 11 linguistic features for each AV sample. These features, identified in the research by Boenninghoff et al. (2019), are crucial for analyzing writing styles in textual content. The linguistic features generated by ChatGPT were recorded as labels for the AV explanations in our dataset.

To ensure a tight coupling between classification labels and linguistic explanations during the data collection phase, we crafted prompts that incorporate information from the classification labels. For example, a prompt might state "*Text1 and Text2 are written by the same author*" to guide the explanation process. A detailed example of the explanation generation prompt is shown in Appendix A.1. These prompts enhance the relevance of the collected explanations, thereby enriching the explanatory capacity of the InstructAV model.

Consistency Verification The congruity between classification labels and linguistic explanations is an important aspect that impacts the trustworthiness of the explanation data collected. Although

[2]https://www.yelp.com/dataset

ChatGPT is proficient in generating credible linguistic explanations, there is an inherent risk of producing linguistic analyses that may not fully align with the given classification labels. To address this challenge and ensure a high degree of alignment between classification labels and linguistic explanations, we have implemented a rigorous verification process.

This verification process scrutinizes the linguistic analysis, focusing on identifying "similarities" or "differences" in the texts. It then evaluates whether these observations are consistent with the classification label assigned to the texts. This step, which forms the second stage of our process, is visualized in Figure 1 as an integral component of the InstructAV framework.

The key to this process is the construction of instruction-tuning data, which serves a dual purpose: facilitating AV classification prediction and aiding in the generation of explanations. An example of the instruction-tuning data is shown in Appendix A.2. This verification step is essential for ensuring that the InstructAV model is not only accurate in its predictions but also capable of generating explanations that are coherent, relevant, and aligned with the classification outcomes, thereby enhancing the overall efficacy and reliability of the InstructAV framework.

3.2 Fine-tuning with LoRA

Adapting LLMs for AV tasks can be a demanding process, particularly due to the significant computational resources and extensive labeled data typically required for fine-tuning. To address this challenge, we have incorporated a PEFT method known as LoRA (Hu et al., 2021) into our approach for adapting LLMs to AV tasks.

LoRA introduces an innovative strategy for updating the parameters of a weight matrix. It achieves this by decomposing the matrix into a product of two low-rank matrices, thereby effectively reducing the number of trainable parameters while maintaining the capability to process high-dimensional matrices. During the model's forward pass, the LoRA matrices compute a rank-decomposed version of the original weights. These recalculated weights are then utilized in the computation of attention or feed-forward activations. Crucially, this method ensures that the core parameters of the LLM remain unchanged, facilitating adaptation to new data distributions or tasks without a complete overhaul of the model.

Dataset	#Train	#Test	Avg length
IMDB	10,000	1,000	303
Twitter	10,000	1,000	16
Yelp	10,000	1,000	154

Table 1: Dataset statistics for IMDB, Twitter and Yelp.

The LoRA architecture is encapsulated in the following equation:

$$h = W_0 x + \Delta W_0 x = W_o x + BAx \qquad (1)$$

where $W_0 \in \mathbb{R}^{d \times d}$ is a pre-trained weight matrix in LLMs, which is associated with two adjustable low-rank matrices, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times d}$, d corresponds to the hidden dimension of the attention layers, and r is the adaptation rank, which is selected such that $r \ll d$. Within the training phase, W_0 remains frozen, exempt from gradient updates, whereas A and B are dynamic, containing the parameters subject to training.

4 Experiments

4.1 Experiment Settings

Datasets. Our evaluation of the proposed InstructAV framework encompasses experiments on three distinct AV datasets: IMDB62 (Seroussi et al., 2014), Twitter (Schwartz et al., 2013), and Yelp Reviews [3]. To facilitate a comprehensive experimental test, we extracted a random subset of 10,000 samples from each dataset to form the training sets, and selected 1,000 samples from each for the test sets. As detailed in Section 3.1, these training samples were utilized to compile instruction-tuning data, which includes linguistic explanations. Table 1 shows the statistics of the datasets.

To evaluate the AV classification and explanation tasks performance, we constructed two distinct types of dataset settings, each incorporating varying levels of information:

1. *Classification*: This dataset involves integrating a question alongside two texts as the input and employs the LoRA method to fine-tune the model for classification tasks. The expected output is a straightforward binary classification indicating whether the two texts are written by the same author, formatted as "*The correct answer is yes/no*".

2. *Classification and Explanation*: In this setting, we augment the classification data with

[3] https://www.yelp.com/dataset

Dataset Setting	Model	IMDB	Twitter	Yelp
Classification	BERT	0.663	0.701	0.622
	DistilBERT	0.527	0.575	0.542
	ALBERT	0.646	0.702	0.603
	InstructAV (LLaMA-1-7B)	0.688	0.588	0.541
	InstructAV (OPT-6.7B)	0.569	0.503	0.525
	InstructAV (LLaMA-2-7B)	**0.915**	**0.734**	**0.632**
Classification & Explanation	PromptAV-2shot (ChatGPT)	0.667	0.645	0.539
	PromptAV-4shot (ChatGPT)	0.635	0.667	0.544
	PromptAV-8shot (ChatGPT)	0.601	0.648	0.564
	InstructAV (LLaMA-1-7B)	0.792	0.619	0.589
	InstructAV (OPT-6.7B)	0.749	0.707	0.561
	InstructAV (LLaMA-2-7B)	**0.938**	**0.747**	**0.661**

Table 2: Performance of InstructAV and the baselines on two dataset settings. Highest Acc are **bolded**.

linguistic analysis to empower the model to generate robust explanations for the AV predictions. The LLMs are fine-tuned to not only predict the classification labels, but also provide an analysis of eleven linguistic features of the two texts. This added layer of analysis offers a reasoned explanation behind the classification decision, thereby enhancing the model's interpretability and reliability.

These two dataset settings enable a thorough investigation into how the addition of explanatory components influences the performance of the classification task, providing insights into the efficacy and versatility of the InstructAV framework.

Baselines. For the *Classification* task, our baseline models comprised BERT (Devlin et al., 2018) and its variants, DistilBERT (Sanh et al., 2019), and AlBERT (Lan et al., 2019). These models were selected for their widespread usage and proven effectiveness in AV classification tasks (Brad et al., 2021; Tyo et al., 2022; Fabien et al., 2020). However, these models are not inherently geared towards text generation tasks.

For tasks involving *Classification and Explanation*, autoregressive models like GPT, which generate text sequences by predicting each word based on the preceding text, are more appropriate. As a baseline for the ChatGPT model, we adopted the PromptAV methodology proposed by Hung et al. (2023). Although prior research indicates that PromptAV may be slightly less effective than BERT in terms of classification accuracy, it offers distinct advantages, including the ability to bypass extensive training and generate explanations.

Consequently, our comparative analysis is designed to evaluate not only the classification accuracy but also the quality of explanations produced by both InstructAV and PromptAV. This comprehensive assessment aims to provide a clear understanding of the strengths and potential limitations of each approach in the context of AV tasks.

Implementation. In our research, we employed three different LLMs to validate the efficacy of the InstructAV framework. These models include OPT-6.7B (Zhang et al., 2022), LLaMA-1-7B (Touvron et al., 2023a), and LLaMA-2-7B (Touvron et al., 2023b). Our experiments were conducted using two NVIDIA A6000 GPUs, encompassing both the fine-tuning and inference phases.

For the fine-tuning process, we utilized the LoRA approach, setting the rank to 8. This configuration introduced an additional 4.1 million parameters, which constitutes only 0.06% of the total 7 billion parameters of the base models. This increase in parameters underscores the effectiveness of LoRA in providing a parameter-efficient fine-tuning approach. The fine-tuning was conducted over 3 epochs using the LLM-Adapters Toolkit (Hu et al., 2023), which is specifically designed for integrating adapters into LLMs.

During the inference phase, all operations were executed deterministically, with a fixed temperature setting of 0.1. This approach ensures consistency and reliability in the model's performance, crucial for assessing the true capabilities of the InstructAV framework. The choice of these diverse LLMs and the strategic use of LoRA for fine-tuning effectively demonstrate the versatility of the InstructAV framework across different model architectures.

4.2 Evaluation Metrics

The evaluation of the InstructAV on AV classification and explanation generation tasks. This dual-focused approach allows for a thorough assessment of the framework's capabilities in key areas of AV.

For AV classification tasks, accuracy serves as the primary metric. This metric effectively

IMDB				
Model	R-1	R-2	R-L	BS
PromptAV-2shot	.379	.147	.227	.844
InstructAV (LLaMA-2)	**.689**	**.434**	**.515**	**.907**
Twitter				
PromptAV-2shot	.445	.203	.287	.856
InstructAV (LLaMA-2)	**.689**	**.420**	**.542**	**.904**
Yelp				
PromptAV-2shot	.377	.142	.243	.856
InstructAV (LLaMA-2)	**.716**	**.429**	**.592**	**.912**

Table 3: Performance of InstructAV and PromptAV on automatic metrics. Higher Acc are **bolded**. R-1, R-2, R-3 and BS represent ROUGE-1, ROUGE-2, ROUGE-3 and Bert_Score, respectively.

IMDB				
Model	Cov	Rel	Rea	Per
ChatGPT	11	4.42	4.12	4.22
PromptAV-2shot	7	3.8	3.84	3.26
InstructAV (LLaMA-2-7B)	11	4.24	3.98	4.03
Twitter				
ChatGPT	11	4.66	4.42	4.65
PromptAV-2shot	7	3.01	3.24	2.94
InstructAV (LLaMA-2-7B)	11	4.48	4.25	4.28
Yelp				
ChatGPT	11	4.56	4.44	4.65
PromptAV-2shot	7	2.87	3.02	3.06
InstructAV (LLaMA-2-7B)	11	4.3	4.32	4.36

Table 4: Performance of InstructAV on human evaluation. Cov means Coverage, Rel means Relevance, Rea means Reasonableness and Per means Persuasiveness

measures the models' proficiency in determining whether two texts have been authored by the same individual. The evaluation of the framework's performance in AV explanation generation tasks necessitates a different approach. Here, the focus shifts towards assessing the quality of the linguistic analyses generated by the models. Unlike accuracy, which is quantitatively measurable, the quality of linguistic explanations presents a more complex evaluation challenge, as it involves subjective elements related to the coherence, relevance, and clarity of the explanations. Given the nuanced nature of evaluating explanation quality, we adopt a two-pronged approach, employing both automatic and human evaluation metrics.

Automatic Evaluation for Explanations. In our automated assessment of explanation quality for both the InstructAV model and baseline models, we measure the similarity between generated and reference explanations, using ChatGPT-generated texts as reference standards. We employ metrics like ROUGE-1, ROUGE-2, and ROUGE-L (evaluating unigram, bigram overlaps, and longest common subsequences, respectively, for content coverage and structural fluency), and BERT_Score (assessing semantic quality). Higher scores across these metrics indicate better explanation quality, ensuring a comprehensive evaluation of the explanations in terms of content accuracy, logical coherence, and contextual relevance.

Human Evaluation for Explanations. To supplement the automated metrics for evaluating explanation quality in the InstructAV framework, we also implemented a human evaluation. This approach involved randomly selecting 100 examples from each of the IMDB, Twitter, and Yelp Reviews datasets for qualitative assessment by human evaluators. These evaluations included not only explanations generated by InstructAV and baseline models, but also those produced by ChatGPT, serving as a high-standard reference. Three linguistic experts were enlisted to independently evaluate each explanation across four criteria:

- **Coverage**: Assessing the extent to which explanations include the expected linguistic features. Evaluators checked the number of features covered in each explanation, with baseline models like PromptAV covering 0 to 7 features, and InstructAV, along with ground truth, extending from 0 to 11 features.
- **Relevance**: Evaluators scored the relevance of the explanations to the original texts on a 5-point Likert scale, ranging from 1 (completely irrelevant) to 5 (perfectly relevant).
- **Reasonableness**: This involved assessing the logical soundness of each feature's explanation, using a 5-point Likert scale, from 1 (very unreasonable) to 5 (completely reasonable).
- **Persuasiveness**: Evaluators rated the persuasiveness of the explanations on a scale from 1 (not persuasive) to 5 (highly persuasive).

4.3 Experiment Results

Classification Results. We evaluate the InstructAV framework for AV tasks, utilizing both classification and explanatory dataset settings. The primary goal is to explore how linguistic analysis can improve the model's performance in AV classification tasks. The results, detailed in Table 2, include the accuracy of InstructAV with various LLMs and baseline models. A key finding is that InstructAV,

particularly when paired with LLaMA-1-7B and LLaMA-2-7B, outperforms baseline models in all datasets when using only classification data. Notably, with LLaMA-2-7B, InstructAV achieves a 25.2% improvement over the highest-performing baseline, BERT, on the IMDB dataset.

It's crucial to note that when using explanatory data, models are tasked with generating both AV classification predictions and linguistic analyses. In our tests, no explanation information was included in the input, ensuring an equitable comparison with the classification-only approach. Our findings show that all InstructAV variations, with different LLMs, greatly benefit from training on explanatory generation, exhibiting significant improvements. Specifically, InstructAV with LLaMA-2-7B registers a 27.1% improvement over PromptAV with 2-shot prompts, which employs ChatGPT (175B) as its foundation. This highlights that training InstructAV to produce both AV classification predictions and linguistic explanations can substantially boost classification accuracy.

Automatic Evaluation Results on Explanations. Our InstructAV framework demonstrates that training in explanation generation can significantly improve AV classification performance. Additionally, we aim to assess whether InstructAV can generate high-quality explanations. For this purpose, we selected PromptAV-2shot and InstructAV with LLaMA-2-7B, the best-performing variation of each framework, and subjected them to both automatic and human evaluations focused on the quality of their generated linguistic explanations.

The results of the automatic evaluation are presented in Table 3. These results reveal that InstructAV consistently surpasses PromptAV across all datasets and evaluation metrics. Notably, the ROUGE-1 and ROUGE-2 scores highlight InstructAV's superior performance in achieving content overlap at both the word and phrase levels. Moreover, the ROUGE-L metric indicates that InstructAV is more proficient in maintaining sentence-level structure and fluency. Furthermore, the BERT_Score supports the observation that the explanations generated by InstructAV are semantically closer to the ground truth. This comprehensive evaluation underscores InstructAV's capability not only in improving AV classification accuracy but also in generating linguistically coherent and contextually relevant explanations.

	IMDB	Twitter	Yelp
Top 25	**0.92**	**0.8**	**0.84**
Bottom 25	0.88	0.68	0.72

Table 5: Performance of InstructAV on Top 25 and Bottom 25 explanation. Higher accuracies are bolded.

Human Evaluation Results on Explanations In comprehensively evaluate the generated explanations and to assess the quality of the linguistic analyses produced by ChatGPT, which serves as our ground truth, we have conducted a human evaluation using four key metrics: Coverage, Relevance, Reasonableness, and Persuasiveness. The results of this human evaluation are presented in Table 4.

Our findings from this human evaluation process show that ChatGPT achieves the highest scores, underscoring its effectiveness in generating high-quality explanations. This result validates our methodological choice of using ChatGPT-generated explanations as both a source for training data and a benchmark for ground truth in our testing scenarios. Importantly, the results also reveal that InstructAV, particularly with the LLaMA-2-7B model, not only surpasses the performance of PromptAV-2shot but also attains a level of explanation quality that is comparable to ChatGPT. This outcome is significant as it demonstrates that InstructAV can produce explanations that are not only accurate but also contextually relevant, logically sound, and convincing to human evaluators. Such a capability is essential for applications where understanding the rationale behind model predictions is as important as the predictions themselves.

Correlation between Explanation and Classification. To explore the relationship between explanation quality and classification accuracy, we have selected two distinct subsets of InstructAV samples: the top 25% with the highest average human evaluation scores and the bottom 25% with the lowest average human evaluation scores. We then calculated the classification accuracy for each of these subsets, and the results are presented in Table 5. Our analysis clearly indicates that samples associated with higher quality explanations consistently achieve superior classification accuracy. This finding underscores the effectiveness of training InstructAV to not only generate AV predictions but also provide linguistic explanations. Such training not only enhances the model's performance in AV classification tasks but also enhances its ability to produce valuable linguistic analyses.

Instruction: *Please decide if the following Text1 and Text2 are written by the same authors. 'yes' means from the same author, 'no' means not from the same author. Then, provide an analysis based on writing styles.*

Text 1: *The British Soap Awards 2005 I am a fan of only two soaps , Coronation Street and EastEnders , but I am always going to watch this programme to see what the other soaps are like , if they deserve to win something , and what I may be missing out on . I always want to agree with the awards ...* [Truncated due to length restriction]

Text 2: *Father of the Bride As far as I can work out , it is after this film that the career of Steve Martin has had very few laughs , which is a shame , he just needs to find the right film . Anyway , this is the remake of the 1950 film , and it sees middle aged father George Banks (Martin) with wife Nina (Diane Keaton) having to prepare the wedding for their 21-year-old daughter Annie Banks (Kimberly Williams-Paisley) ...* [Truncated due to length restriction]

Generated Output

Classification: The correct answer is yes.
Explanation: Upon analyzing Text 1 and Text 2 based on the listed writing style characteristics, we find the following similarities that could suggest that both texts were written by the same author:
Writing Style: Both texts provide personal opinions and evaluations of the subject matter. Text 1 focuses on the British Soap Awards, while Text 2 provides a review of the film \"Father of the Bride.\""
Expressions and Idioms: Both texts use colloquial expressions and idioms, such as \"if they deserve to win something\" in Text 1 and \"he just needs to find the right film\" in Text 2.
Tone and Mood: Both texts have an informal and conversational tone. The author expresses their personal preferences and opinions in both texts.
Sentence Structure and Syntax: Both texts use a mix of compound and simple sentences. For example, \"I always want to agree with the awards that are given according to the clips/photos shown\" in Text 1 and \"Martin does do well, Keaton offers a good support, and Short adds giggles as the foreign wedding organiser\" in Text 2.
Punctuation Style: Both texts use consistent punctuation, including commas and full stops.
Special Characters Style, Capitalization Style: Both texts capitalize proper nouns and show titles. However, Text 2 uses unconventional capitalization for emphasis, such as \"Very good!\"
Compound and Separate Spelling: No significant differences in spelling patterns are observed between the two texts.
Acronyms and Abbreviations: Both texts use acronyms and abbreviations, such as \"EE\" for \"EastEnders\" in Text 1 and \"B.D.Wong\" for \"B.D. Wong\" in Text 2.
Characters Style: Both texts mention specific actors and their roles, indicating a familiarity with the subject matter.
Diatopic Variations and Foreign Languages: Both texts appear to be written in standard English without diatopic variations or foreign language references.
Other Relevant Aspects: Both texts provide specific details about the subject matter, indicating a personal familiarity or interest in the topic. The author's opinions and evaluations are evident in both texts.

Table 6: A text pair example with the generated responses from InstructAV

5 Case Study

We have included randomly selected examples to showcase the capabilities of InstructAV in generating both classification predictions and language feature-based explanations. One such example from the IMDB dataset is detailed in Table 6. This example demonstrates how InstructAV first provides a classification prediction and then follows it with detailed explanations anchored in specific language features pertinent to the classification.

These examples highlight InstructAV's proficiency in delivering accurate classifications paired with clear and credible explanations. Notably, the explanations are tailored to each specific sample, rather than being generic or broad. This customization enhances the explanations' effectiveness, as they are directly correlated with the content of the original text. For instance, the model explicates features such as writing style and use of expressions by incorporating words and phrases directly from the text in question. This approach not only bolsters the relevance of the explanations but also their trustworthiness, providing users with insightful and contextually grounded rationale behind each classification decision made by InstructAV.

6 Conclusion

This research presents InstructAV, an innovative approach to AV that leverages LLMs with a PEFT method. Our study establishes InstructAV as a significant advancement in the AV domain, showcasing its ability to enhance classification accuracy and provide clear and coherent explanations for its decisions. The contributions of this paper, including the development of the InstructAV framework, the creation of three instruction-tuning datasets with reliable linguistic explanations, and the demonstration of the framework's effectiveness through both automated and human evaluations, mark a pivotal step forward in AV research. InstructAV's balanced focus on accuracy and explainability positions it as a state-of-the-art AV solution.

7 Limitations

The InstructAV framework, while achieving state-of-the-art performance in AV classification and explanation tasks, does face a notable limitation in its current form. Specifically, when tasked with generating both AV classification predictions and linguistic explanations, InstructAV is constrained to produce a maximum of 512 new tokens. This restriction can lead to significantly longer inference times. For example, generating explanations for 1,000 samples in a long-text dataset like IMDB can take around 4-5 days, while in shorter-text datasets such as Twitter, this time is reduced to approximately 2 days. Addressing this challenge, future work will focus on developing methods to enhance the efficiency of the inference process within InstructAV, particularly in the context of generating explanations.

References

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing unmasking for short texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 654–659.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE.

Florin Brad, Andrei Manolache, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. 2021. Rethinking the authorship verification experimental setups. *arXiv preprint arXiv:2112.05125*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Oren Halvani, Christian Winter, and Lukas Graner. 2019. Assessing the applicability of authorship verification methods. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–10.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Ka-Wei Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. *arXiv preprint arXiv:2310.08123*.

Keenan Jones, Jason RC Nurse, and Shujun Li. 2022. Are you robert or roberta? deceiving online authorship attribution models using neural text generators. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 429–440.

Stefanos Konstantinou, A Zinonos, and J Li. 2022. Different encoding approaches for authorship verification. CLEF.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*, 70:101241.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on empirical methods in natural language processing*, pages 1880–1891.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2011. Authorship attribution with latent dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 181–189.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2014. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Efstathios Stamatatos. 2016. Authorship verification: a review of recent advances. *Research in Computing Science*, 123:9–25.

Yitao Sun, Svetlana Afanaseva, and Kailash Patil. 2023. Stylometric and neural features combined deep bayesian classifier for authorship verification. *Working Notes of CLEF*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*.

Andric Valdez-Valenzuela, Jorge Alfonso Martinez-Galicia, and Helena Gómez-Adorno. 2023. Heterogeneous-graph convolutional network for authorship verification.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wanwan Zheng and Mingzhe Jin. 2023. A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1584.

A Appendix

A.1 Prompt Templates

A.2 Collected Datasets

Task: Text1 and Text2 are written by the same author. Please analyze their writing styles and explain why they are written by the same author/ different authors. You can refer to the following characteristics of writing style. 1. writing style. 2. expressions and Idioms. 3. tone and mood. 4. sentence structure and syntax. 5. punctuation style. 6. special characters style, capitalization style. 7. compound and separate spelling. 8. acronyms and abbreviations. 9. characters style. 10. Diatopic variations and foreign languages. 11. any other relevant aspect.

Text 1: *The British Soap Awards 2005 I am a fan of only two soaps , Coronation Street and EastEnders , but I am always going to watch this programme to see what the other soaps are like , if they deserve to win something , and what I may be missing out on. I always want to agree with the awards that are given according to the clips / photos shown , and that is what I also enjoy , watching back on the past year ...* [Truncated due to length restriction]

Text 2: *Father of the Bride As far as I can work out , it is after this film that the career of Steve Martin has had very few laughs , which is a shame , he just needs to find the right film. Anyway , this is the remake of the 1950 film , and it sees middle aged father George Banks (Martin) with wife Nina (Diane Keaton) having to prepare the wedding for their 21-year-old daughter Annie Banks (Kimberly Williams-Paisley). They had never met her fianc0̆0e9e Bryan MacKenzie (George Newbern) , living an upper class with his parents in Bel-Air , but George tries to keep his feelings about wedding preparations and this new man in his daughter's life intact ...* [Truncated due to length restriction]

Please follow the format of the analysis method in the demostrations. You will be given 2 demostrations.
Demostration Start:
Text 1: Mama's homemade remedy for laughs! Mama Harper develops her mother's concoction to help her family get over an illness like the flu. After her son and bimbo-in-law Naomi who is now expecting their child realized that they can't afford it. They try a get rich scheme which falls on the unmentioned ingredient , alcohol.....[Truncated due to length restriction]
Text 2:Not so Bad! Remember that this show was one of the first sitcoms in syndication in the 1980s. It was a new thing. When a sitcom or drama could not get a network , they went with syndication. I thought this show was alright in syndication because it reached a wide audience. I often saw it on Saturday evenings usually before Mama's Family. I was a kid so I didn't know better.....[Truncated due to length restriction]

Upon analyzing Text 1 and Text 2 based on the listed writing style characteristics, we find the following similarities that could suggest that both texts were written by the same author:
Writing Style: Both texts seem to be reminiscing about older television shows. The author uses a mixture of narrative and opinion in both texts, indicating a personal connection or memory attached to the subject matter.
Expressions and Idioms: The expressions "I miss this show" and "I don't recall ever not watching it" showcase a personal touch and sentimentality in both texts.
Tone and Mood: Both texts have a nostalgic and somewhat informal tone. The author shares memories and impressions about shows from the past, using phrases like "I was a kid so I didn't know better" and "Still, I miss this show."
Sentence Structure and Syntax: Both texts employ a mix of compound and simple sentences. For instance, "When a sitcom or drama could not get a network, they went with syndication" and "Mama Harper develops her mother's concoction to help her family get over an illness like the flu."
Punctuation Style: Both texts employ spaces before and after commas inconsistently, such as in "$180 , 000" and "Marla Pennington , Edie McClurg."
Special Characters Style, Capitalization Style: Both texts capitalize show titles like "Mama's Family" and show proper nouns. However, both texts sometimes lack proper capitalization at the start of sentences, such as "Still" in Text 1.
Compound and Separate Spelling: No significant differences in spelling patterns are observed between the two texts. Acronyms and Abbreviations: The second text uses "1980s" to denote the decade, showcasing a preference for this kind of abbreviation.
Characters Style: The character portrayal in both texts has a casual undertone. For example, "bimbo-in-law Naomi" and "the guy who played her father."
Diatopic Variations and Foreign Languages: Both texts appear to be written in American English and reference American television shows.
Other Relevant Aspects:Both texts contain anecdotal or subjective experiences like "I was a kid so I didn't know better" and "Still, I miss this show." The author provides specific details about show characters, indicating a personal familiarity with the subject. There's a recurring theme of syndication and older television programming in both texts, further suggesting the same author might have written them.
In conclusion, based on the aforementioned similarities in writing style, expressions, tone, and other characteristics, it is plausible that Text 1 and Text 2 were written by the same author.

Table 7: Prompt used for generating explanation labels

Classification Datasets
Instruction: Please decide if the following Text 1 and Text 2 are written by the same authors. 'yes' means from the same author, 'no' means not from the same author.

Text 1: *Send in the Clowns!!! I don't know why we haven't heard of Judy Cohen since this episode. She can break your heart singing Stephen Sondheim's Send in the Clowns. She plays herself as an aspiring singer who beats herself down very much because she is overweight and not the classic beauty.....*[Truncated due to length restriction]
Text 2: *Could have been better! Columbo's Most Dangerous Match is about chess. An unlikely subject or motive for murder in my opinion. The star is Manchurian Candidate's Lawrence Harvey who plays a deaf chess champion.....*[Truncated due to length restriction]

Output: The correct answer is yes.

Classification & Explanation Datasets

Instruction: *Please decide if the following Text 1 and Text 2 are written by the same authors. 'yes' means from the same author, 'no' means not from the same author. Then, provide an analysis based on writing styles.*

Text 1: *Send in the Clowns!!! I don't know why we haven't heard of Judy Cohen since this episode. She can break your heart singing Stephen Sondheim's Send in the Clowns. She plays herself as an aspiring singer who beats herself down very much because she is overweight and not the classic beauty.....*[Truncated due to length restriction]
Text 2: *Could have been better! Columbo's Most Dangerous Match is about chess. An unlikely subject or motive for murder in my opinion. The star is Manchurian Candidate's Lawrence Harvey who plays a deaf chess champion.....*[Truncated due to length restriction]

Output: The correct answer is yes. Analyzing Text 1 and Text 2 based on the listed writing style characteristics, we find the following similarities that could suggest that both texts were written by the same author:
Writing Style: Both texts provide detailed commentary and analysis of specific episodes of a TV show. The author shares personal opinions and emotional responses to the episodes.
Expressions and Idioms: Both texts use phrases like "break your heart" and "a hell of a game" to evoke emotional reactions and convey personal opinions.
Tone and Mood: Both texts exhibit enthusiasm and engagement with the subject matter. The author expresses admiration for certain performances and emphasizes their impact.
Sentence Structure and Syntax: Both texts use a mix of simple and compound sentences. The sentences focus on describing specific events and interactions between characters.
Punctuation Style: Both texts use consistent punctuation, including proper use of commas and ellipses. They enhance the flow and expressiveness of the narrative.
Special Characters Style, Capitalization Style: Both texts consistently capitalize proper nouns and show titles. The author pays attention to the details and presents them accurately.
Compound and Separate Spelling: Both texts use standard spelling patterns without significant variations.
Acronyms and Abbreviations: Neither text employs acronyms or abbreviations.
Characters Style: Both texts describe specific characters and their roles within the TV show episodes. The author demonstrates familiarity with the characters' interactions and motivations.
Diatopic Variations and Foreign Languages: Both texts are written in English without reference to specific geographical locations or foreign languages.
Other Relevant Aspects: Both texts contain personal appeals to specific individuals, such as "Where are you Judy Cohen?" and "You almost feel sorry for him." The author shows a deep connection to the subject matter and a desire for continued engagement.
In conclusion, based on the aforementioned similarities in writing style, expressions, tone, and other characteristics, it is likely that Text 1 and Text 2 were written by the same author.

Table 8: Example of collected **Classification** Datasets and **Classification & Explanation** Datasets