# Blue Sky: Multilingual, Multimodal Domain Independent Deception Detection*

Dainis Boumber
dboumber@uh.edu

Rakesh M. Verma
rmverma2@Central.UH.EDU

Fatima Zahra Qachfar
fqachfar@uh.edu

**Abstract**

Deception, a pervasive aspect of communication, has undergone a significant transformation in the digital age. With the globalization of online interactions, individuals are communicating in multiple languages, mixing languages on social media. A variety of data is now available in many languages, while the techniques for detecting deception are similar across the board. Recent studies have shown the possibility of the existence of universal linguistic cues to deception across domains within the English language; however, the existence of such cues in other languages remains unknown. Furthermore, the practical task of deception detection in low-resource languages is not a well-studied problem due to the lack of labeled data. Another dimension of deception is multimodality. For example, in fake news or disinformation, there may be a picture with an altered caption. This paper calls for a comprehensive investigation into the complexities of deceptive language across linguistic boundaries and modalities, and raises the possibility of use of multilingual transformer models and labeled data in a variety of languages to universally address the task of deception detection.

## 1 Introduction

Deception is a complex and pervasive phenomenon with profound implications for various domains, including security, law enforcement, healthcare, and human-computer interaction. The ability to accurately identify deception has long been a critical goal for researchers and practitioners alike. Traditional methods for deception detection (DD) have primarily relied on linguistic cues and textual analysis [24, 26]. A DD task is typically a binary classification problem, aiming to label a statement as being deceptive or not. Less often, the goal is to categorize a statement as falling into one of the more or less deceptive categories. It is a problem of growing importance that is made more challenging by the need to build different datasets and detectors for the ever-increasing variety of domains and tasks where deceptive language poses a threat. However, these methods often fall short in the face of sophisticated deceivers who can manipulate language effectively, leaving the task of deception detection far from foolproof [16]. Recently, there has been a paradigm shift towards more comprehensive and robust approaches to deception detection, which leverage multimodal data sources. This shift recognizes that deception is not confined to language alone and that individuals may convey deceptive information through various channels, including speech, facial expressions, body language, and by using different languages[1] Another frequently debated topic is the transfer of linguistic cues of deception across domains and modalities. The need for domain-independence in deception detection is paramount, since there are many manifestations of deception.

We call for a holistic approach to deception detection, focusing on the integration of multimodal (multiple modes of communication) and multilingual (cross-linguistic) data, while maintaining domain independence. We propose leveraging cutting-edge advances in natural language processing (NLP), computer vision, and machine learning (ML) to enhance the accuracy and robustness of deception detection across a wide array of applications and settings. The research in this area must aim to address several critical challenges in deception detection, including the integration of non-verbal cues from multiple modalities (e.g., speech, facial expressions, gestures, image/video attachments) and the consideration of linguistic variations across different languages. By developing a domain-independent approach, we call for the creation of a versatile approach that can be applied to diverse scenarios, from border security and criminal investigations to healthcare diagnostics and online content moderation.

In this paper, we will present the theoretical foundations of multimodal multilingual deception detection through existing work and suggest the methodologies to be employed. The goal of this blue sky paper is to open a new direction of research that will usher forth valuable insights into a comprehensive approach to deception detection that transcends linguistic and contextual boundaries, opening up new possibilities for enhancing trust, security, and decision-making across various domains.

## 2 Why is it a Blue Sky Idea? Why Now?

The proliferation of deceptive attacks such as fake news, phishing, and disinformation is rapidly eroding trust in Internet-dependent societies. Social-media platforms have come under severe scrutiny regarding how they police con-

*All authors are with the University of Houston

[1] However, multilingual deception detection efforts are relatively fewer.

tent. Facebook and Google are partnering with independent fact-checking organizations that typically employ manual fact-checkers. With the advent of Large Language Models, such as ChatGPT, things are only going to get worse.

Building single-domain detectors is sub-optimal becayse it requires time and ultimately means one can only react to new forms of deception **after** they emerge. Our goal here is to spur research on *domain-independent* deception. Unfortunately, research in this area is currently hampered by the lack of computational definitions and taxonomy, high-quality datasets, and systematic approaches to domain-independent deception detection. Thus, results are neither generalizable nor reliable, leading to much confusion.

## 3 Related Work

In the past few years, there have been several studies of applying computational methods to deal with deception detection in a single domain. For fake news, [3] used topic models and [10] used Bag of Words (BoW) and BERT [4] embedding. State-of-the-art (SOTA) in phishing detection has been dominated by classical supervised machine learning approaches and deep neural nets [6]. More recently, word embeddings produced by BERT [4], a character-level CNN, and sentence embeddings from Sentence-BERT (SBERT) [18], were used to find emails exhibiting psychological traits most dominant in phishing texts [22]. In detection of opinion spam and fake reviews, weakly supervised graph networks have been recently used with some success [15]. [7, 17, 14] used part-of-speech tags and context-free grammar parse trees, behavioral features, and spatial-temporal features, respectively. Neural network methods for spam detection consider the reviews as input without specific feature extraction.In [19], authors used a gated recurrent neural network to study the contextual information of review sentences. DRI-RCNN [27] used a recurrent network for learning the contextual information of the words in the reviews. Several studies on cross-domain deception detection have been published, as well [11, 20, 21]. Recently, a quality domain-independent deception dataset was introduced in [26], with the empirical evidence suggesting that large language models such as BERT and RoBERTa perform well on individual tasks when fine-tuned on a combination of out of domain deceptive texts. Closer to our stated goals, [25] created a multi-modal deception detection tool that used early deep learning models and word embeddings, although ultimately the performance was not always robust and it lacked domain-independence capabilities. Finally, [8] propose a framework for evaluating the robustness of deception detection models across two domains (Twitter and Reddit), modalities (Text, images), and five languages.

**Datasets** To create a diverse and versatile dataset for training AI models in multilingual and multimodal deception detection, we recommend utilizing a range of deception data sources provided in [2]. This approach will help ensure that models are not limited to a single language or modality, and can effectively detect deception across different cultures and communication channels.

## 4 Challenges and Opportunities

Here, we discuss the major research challenges and research opportunities. One of the main problems that makes this a blue-sky idea is the fact that there is no consensus regarding the transferability of deceptive cues across domains even within a single modality. For example, a recent review of text mode deception literature [9] found unclear and contradictory results and concluded there was no evidence of deception's stylistic trace. However, a more recent publication [26] presented evidence to the contrary insofar as DD within the text. Other substantial challenges include:

1. *Defining deception computationally.* So far, deception has been defined using the intent of the deceiver, but the attacker is elusive in the real world, so intentions are impossible to access.

2. Giving a taxonomy for deception that is comprehensive and useful to guide further research. For example, the taxonomy should help in building a quality general deception dataset and then generalized deception detection models. Ensuring that it is high quality is a challenge also.

3. Finding a common basis for the different forms of deception.

4. Finding common cues and invariants across the different forms of deception.

5. *Dealing with imbalanced data.* Deceptive attacks, by their nature, would be either targeted, e.g., spearphishing, or overly broad such as spam or phishing.

6. *Distributed nature.* People and companies are not comfortable sharing sensitive information such as targeted attacks (spearphishing). Can we design models that can work with limited shared data.

7. *Human in the loop.* Can the detector improve human ability? Can the human improve the detector ability with just a few examples? Or by providing access to his/her cognitive load through a sensor?

## 5 Defining Success

Ideally, we would achieve a deception detection model that does not need any labeled data to detect new forms of attacks since it is based on invariants of deception. However, this may be too difficult a holy grail to achieve. Thus, success would be a detector that helps the human in the

loop do significantly better at resisting attacks (e.g., a novice email user is able to detect quite sophisticated spearphishing attacks). If we are able to achieve this goal, then we can start researching the problem of building teachable detectors so that the human and the detector can improve each other.

## 6 A Possible Solution

In this section, we present a preliminary solution to address the Blue Sky Problem outlined in this work. Our proposed approach leverages the power of advanced, large-scale, multilingual, and multimodal contextual learners, complemented by Retrieval Augmented Generation (RAG). It's important to note that this approach is just one of several promising avenues warranting further exploration.

To lay the foundation for a potential solution, we initially focus on a single modality, namely text. In this configuration, our objective is to create a system capable of identifying deceptive text in a domain-agnostic and multilingual context. To ensure the system's relevance and effectiveness, we aim to imbue it with additional desirable attributes, including result explainability and robust zero to few-shot performance. To achieve these characteristics, a logical system design might involve the integration of an exceptionally large-scale language model that excels in contextual learning, such as Mistral [12]. This model exhibits the ability to achieve superior performance on the task, irrespective of the domain, and remains resilient in the face of diverse data distributions. To further enhance its explainability, the model can undergo instruction tuning, enabling it not only to deliver answers but also to elucidate its underlying reasoning.

To elevate its already outstanding zero to few-shot capabilities and achieving performance parity with fully fine-tuned specialized models in specific domains, it is advisable to augment the learner with a Retrieval Augmented Generation (RAG) infrastructure, as proposed by [13].

The high-level system design scheme of a standard RAG system is shown in Figure 1. It consists of mining-derived datasets indexed and stored in a vector database, accessed through FAISS [5] (Facebook AI Similarity Search), or any other approximate nearest neighbors algorithm optimized for searching large vector spaces. Combined with Retrieval-Augmented Generation (RAG) techniques [13], this infrastructure delivers critical context to enhance Mistral's performance. This system design is a decision-making tool detecting deceptive text followed with a clear explanation for this decision. Moreover, enhancing LLM input prompt with retrieved context guarantees that the model has all the necessary information to generate a comprehensive response.

When presented with a single instance of a task, typically a query from a user, the steps in solving the problem (each step is referenced in Figure 1 as (1),(2),(3), or (4)) are:

1. *Create Initial Prompt:* Starting with the user query.
2. *Augment Prompt with Retrieved Context:* Merges the initial prompt with the context retrieved from the Vector Store, creating an enriched input for the LLM.
3. *Send Augmented Prompt to LLM:* The LLM receives the enhanced prompt.
4. *Receive LLM's Response:* After processing the augmented prompt, the LLM generates its response.

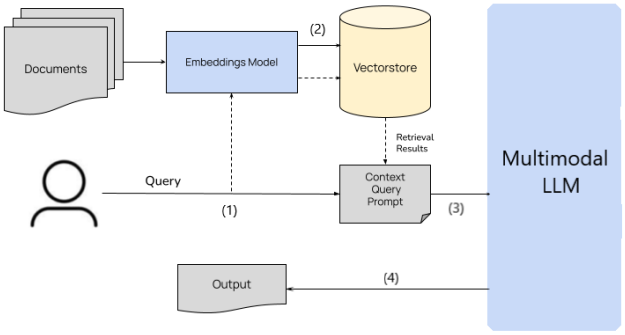


Figure 1: High-level overview of a possible solution using RAG and a multi-modal in-context learner. Dashed line depicts the retrieval of context and its integration into the query; query blocks until it is complete.

To extend this approach to multiple modalities, a multimodal LLM is necessary, e.g., a unified multimodal model, or UnIVAL [23], which unifies text, images, video, and audio into a single model. Thus, it may be possible to use UnIVAL in place of a single-mode model like Mistral [12]. Furthermore, other components, such as RAG, may need to be adjusted as needed. However, a majority of problems to be solved are in the engineering space.

From a research perspective, the main challenge in designing and implementing a functional system as described is model performance as a function of model size and current lack of multimodal in-context learners that are large enough to perform in a satisfactory manner. For text, we notice that models of 2-3B parameters score 0.25 or so on HuggingFace LLM benchmark, 7B score 0.73, and 170B score 0.75; in the other words, there is a huge jump from 3B to 7B parameters followed by a plateau. So to have a truly intelligent model it needs to be at least 7B for a single modality, as a rule of thumb - although it may be possible to bring this number down through quantization and other means. Multiple modalities may require larger models for similar performance. Currently, UnIVAL has only 0.25B parameters. Therefore, we hypothesize that such a system as we described would be made possible by an advance in models like UnIVAL - multimodal transformers that learn in-context and have billions of parameters. Perhaps, Flamingo with 80B parameters [1] can help here.

## 7 Conclusions

In this Blue Sky paper, we introduce a new concept "Multilingual, Multimodal Domain-independence Deception Detection" that unifies diverse investigations, creating a new paradigm for detecting deceitful behavior across languages and modalities. This innovative approach harmoniously connects previous research in the realms of multimodal and cross-lingual deception detection, paving the way for future breakthroughs. We discuss the research challenges related to this concept, and also potential solutions.

**Acknowledgments.**

Research partly supported by NSF grants 2210198 and 2244279, and ARO grants W911NF-20-1-0254 and W911NF-23-1-0191. Verma is the founder of Everest Cyber Security and Analytics, Inc.

## References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, and I. B. et al., *Flamingo: a visual language model for few-shot learning*, 2022.

[2] D. Boumber, R. M. Verma, and F. Z. Qachfar, *A roadmap for multilingual, multimodal domain independent deception detection*, (2024).

[3] W. Ceron, M.-F. de Lima-Santos, and M. G. Quiles, *Fake news agenda in the era of covid-19: Identifying trends through fact-checking content. online social networks and media, 21, 100116*, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.

[5] M. Douze, A. Guzhva, C. Deng, J. Johnson, and G. S. et al., *The faiss library*, arXiv preprint arXiv:2401.08281, (2024).

[6] A. El Aassal, S. Baki, A. Das, and R. M. Verma, *An in-depth benchmarking and evaluation of phishing detection research for security needs*, IEEE Access, 8 (2020), pp. 22170–22192.

[7] S. Feng, R. Banerjee, and Y. Choi, *Syntactic stylometry for deception detection*, in Annual Meeting of the ACL, 2012.

[8] M. Glenski, E. Ayton, R. Cosbey, D. Arendt, and S. Volkova, *Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages*, in Proc. 3rd Int'l Workshop on Rumours and Deception in Social Media (RDSM), Barcelona, Spain (Online), Dec. 2020, ACL, pp. 1–13.

[9] T. Gröndahl and N. Asokan, *Text analysis in adversarial settings: Does deception leave a stylistic trace?*, ACM Comput. Surv., 52 (2019).

[10] A. Hamid, N. Shiekh, N. Said, K. Ahmad, and G. et al., *Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case*, arXiv preprint arXiv:2012.07517, (2020).

[11] Á. Hernández-Castañeda, H. Calvo, A. Gelbukh, and J. J. G. Flores, *Cross-domain deception detection using support vector networks*, Soft Computing, 21 (2017), pp. 585–595.

[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, and D. S. C. et al., *Mistral 7b*, 2023.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, and V. K. et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021.

[14] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, *Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns*, in Proc. ICWSM, vol. 9, 2015, pp. 634–637.

[15] J. Li, L. Yang, and P. Zhang, *Shooting review spam with a weakly supervised approach and a sentiment-distribution-oriented method*, Applied Intelligence, 53 (2022), pp. 10789–10799.

[16] P. Mehdi Gholampour and R. M. Verma, *Adversarial robustness of phishing email detection models*, in Proc. of the 9th ACM Int'l Workshop on Security and Privacy Analytics (IWSPA), 2023, pp. 67–76.

[17] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, *What yelp fake review filter might be doing?*, Proc. ICWSM, (2013).

[18] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019.

[19] Y. Ren and D. Ji, *Neural networks for deceptive opinion spam detection: An empirical study*, Information Sciences, 385-386 (2017), pp. 213–224.

[20] R. Rill-García, L. Villaseñor-Pineda, V. Reyes-Meza, and H. J. Escalante, *From text to speech: A multimodal cross-domain approach for deception detection*, in CVAUI/IWCF/MIPPSNA@ICPR, 2018.

[21] J. Sánchez-Junquera, L. Villaseñor-Pineda, M. M. y Gómez, P. Rosso, and E. Stamatatos, *Masking domain-specific information for cross-domain deception detection*, Pattern Recognit. Lett., 135 (2020), pp. 122–130.

[22] S. Shahriar, A. Mukherjee, and O. Gnawali, *Improving phishing detection via psychological trait scoring*, 2022.

[23] M. Shukor, C. Dancette, A. Rame, and M. Cord, *Unival: Unified model for image, video, audio and language tasks*, Transactions on Machine Learning Research Journal, (2023).

[24] R. M. Verma, N. Dershowitz, V. Zeng, and X. Liu, *Domain-independent deception: Definition, taxonomy and the linguistic cues debate*, Arxiv, (2022).

[25] S. Volkova, E. Ayton, D. L. Arendt, Z. Huang, and B. Hutchinson, *Explaining multimodal deceptive news prediction models*, in International Conference on Web and Social Media, 2019.

[26] V. Zeng, X. Liu, and R. M. Verma, *Does deception leave a content independent stylistic trace?*, in Proc. 12th ACM Conf. on Data and Application Security and Privacy, CODASPY '22, New York, NY, USA, 2022, ACM, p. 349–351.

[27] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, *Dri-rcnn: An approach to deceptive review identification using recurrent convolutional neural network*, Inf. Process. Manag., 54 (2018), pp. 576–592.