CAPITAL UNIVERSITY OF SCIENCE AND

TECHNOLOGY, ISLAMABAD

CogFusionNet: A Cognitive Inspired Fusion Network for Fake News Detection in Multimodal and Bilingual Context for English and Urdu

by

Yasir Hussain

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2026

CERTIFICATE OF APPROVAL

CogFusionNet: A Cognitive Inspired Fusion Network for Fake News Detection in Multimodal and Bilingual Context for English and Urdu

by

Yasir Hussain

(MCS243010)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Rabeeh Ayaz Abbasi	QAU, Islamabad
(b)	Internal Examiner	Dr. Aamir Nadeem	CUST, Islamabad

Dr. Farah Haneef
Thesis Supervisor
May, 2026

Dr. Muhammad Masroor Ahmed
Head
Dept. of Computer Science
May, 2026

Dr. M. Abdul Qadir
Dean
Faculty of Computing
May, 2026

Author's Declaration

I, **Yasir Hussain** hereby state that my MS thesis titled "**CogFusionNet: A Cognitive Inspired Fusion Network for Fake News Detection in Multimodal and Bilingual Context for English and Urdu**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Yasir Hussain)

Registration No: MCS243010

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled " **CogFusion-Net: A Cognitive Inspired Fusion Network for Fake News Detection in Multimodal and Bilingual Context for English and Urdu**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Yasir Hussain)

Registration No: MCS243010

Acknowledgement

"And whoever puts all his trust in Allah, He will be enough for him." Al-Quran [65:1].

I would like to say Alhamdulillah for everything Allah has blessed me with that enabled me to reach here. I would like to express my gratitude to my supervisor Dr. Farrah Haneef who guided me and helped me with her valuable suggestions throughout this work and gave her valuable time, May Allah keep her in his blessings. I am thankful to my parents for their love, prayers, support and everything what I needed. And to my beloved wife and my respected teachers who keep on pushing me for higher education.

Lastly, I would also like to thank everyone who has helped me along the way. Special thanks to Dr. Sabeen Masood for increasing my knowledge and helping me with technical aspects giving motivational support throughout my research journey.

(Yasir Hussain)

Abstract

Digital media is now experiencing a rapidly growing problem with fake news, especially when found in multiple languages and/or low-resource contexts where automated verification is very difficult. Fake news will often use a textual narrative that provides false information combined with an image that is either not directly related to the narrative or loosely related; this means that unimodal methods cannot effectively detect fake news. Existing multimodal methods of detecting fake news typically employ a shallow fusion method of combining different modalities and provide little to no interpretability, thereby limiting their ability to be used in real-world applications. Thus, in attempting to address the above-mentioned problems, this paper presents a new cognitive-inspired, multimodal fake news detection system (CogFusionNet) designed to work in bilingual and multilingual environments, including English, Urdu, and other mixed-language news articles. Using transformer-based multilingual text encoders and a vision transformer, CogFusionNet extracts rich semantic and visual representations from news articles and their accompanying images. An attention-based interaction mechanism has been introduced to model explicitly corroboration of and inconsistencies between text and visual modalities, allowing for one form of human reasoning to take place in validating whether or not the news is real or fake. The performance of this technique is evaluated against the ISOT English Fake News Dataset and the Ax-to-Grind Urdu Fake News Dataset, and additional evaluation on cross-language datasets helps determine how robust this technique is when applied in different languages. The experiments were conducted with 5-fold cross-validation, and the results indicate that CogFusionNet performed very well; it achieved accuracies of 91.5% on English, 89.6% on Urdu, and 90.5% on bilingual datasets, which outperformed many current multilayered and multilingual techniques for detecting fake news (e.g., KGAlign, MAGIC, MCOT, CroMe, and MIMoE-FND). An extensive evaluation using precision, recall and F1-score supports the claim that CogFusionNet performs better than other approaches and does so equivalently across all tested languages. Additionally, ablation studies that compare the loss in performance of the model, due to removing individual component(s)

of the cognitive fusion module, verify its efficacy. The model also demonstrates a significant level of interpretability through the visual representation of attention weights using attention heatmaps, and through human evaluations by bilingual evaluators where the output of the CogFusionNet is in agreement with the evaluators' judgments with an 82% similarity. These results indicate that the proposed framework possesses a high level of robustness, generalizability, and explainability which collectively establish CogFusionNet as a credible and interpretable solution for the detection of fake news across multiple languages and modalities.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ATG	Ax-to-Grind
BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
BiL-FaND	Bilingual Fake News Detection
CLIP	Contrastive Language–Image Pretraining
CLS	Classification Token
CNN	Convolutional Neural Network
CogFusionNet	Cognitive Fusion Network (Proposed Model)
COVID-19	Coronavirus Disease 2019
CSV	Comma-Separated Values
CV	Cross-Validation
F1	F1-Score (Harmonic mean of precision and recall)
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
HEC	Higher Education Commission
ISOT	Information Sciences Institute of Technology
KGAlign	Knowledge Graph Alignment
LLM	Large Language Model
MAGIC	Multimodal Adaptive Graph-based Intelligent Classification
ML	Machine Learning
MNFD	Multimodal Fake News Detection

MSA	Multi-Head Self-Attention
NLP	Natural Language Processing
OCR	Optical Character Recognition
PEFT	Parameter-Efficient Fine-Tuning
PSO	Particle Swarm Optimization
ReLU	Rectified Linear Unit
RQ	Research Question
TN	True Negative
TP	True Positive
ViT	Vision Transformer

Symbols

I	Input image
H, W	Image height and width
C	Number of image channels
P	Patch size
N	Number of image patches
x_i	Flattened image patch
D	Embedding dimension
z_i	Patch embedding
p_i	Positional embedding
Q, K, V	Query, Key, Value matrices
d	Shared latent space dimension
t_i	Text embedding
v_j	Visual embedding
$\hat{y}$	Predicted class label
D_{train}	Training dataset
D_{val}	Validation dataset
D_{test}	Test dataset
F_i	i-th fold in K-Fold cross-validation
K	Number of folds
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
y_i	True class label

Chapter 1

Introduction

Over the last several years, methods used for creating, distributing, and consuming content have dramatically changed. The majority of individuals worldwide rely on social media sites along with digital news websites, and online forums as primary sources of information. These platforms allow users access to real time updates about breaking stories, instant ability to post their thoughts, and actively engage in the discussion of issues affecting their communities at large. Although the overall digital transformation is improving access to information as well as democratizing the creation of content, there are many serious problems posed by the lack of credibility and trust regarding information on these platforms [1].

Perpetrators of misinformation are able to rapidly create and distribute large amounts of false or misleading information designed for maximum impact on emotion, politics, or society within seconds of timing on these platforms. Unlike traditional forms of misinformation, modern fake news is designed to appeal more to your senses instead of rational thought as well as exploiting the emotions of those susceptible to it to create greater likelihood of its persuasiveness and therefore making it nearly impossible for any one person to determine whether something is legitimate or not. We have seen examples of how accurately or inaccurately people use information in making decisions regarding issues such as public health, governmental integrity, social cohesion, and faith in institutions. Significant events, such as health emergencies, political campaigns and the rise of

social movements created significant influence upon how large numbers of people perceive issues based upon the use of misinformation against them. The increase in fake news has prompted the research community to pay special attention to automated detection of false information, especially since early detection methods were manually developed using linguistic features, metadata analysis, and conventional machine learning. As deep learning research progressed, more intelligent models began evolving from these earlier models because of the capabilities of neural networks to identify both semantic and syntactic rules present in the text. Moreover, transformer-based architectures have changed how text analyses are done as they allow for contextual understanding of a language through self-attention mechanisms. They have proven to be extremely effective at detecting fake information in English-language news articles or social media postings [2].

Although technological advances have been made in this area of study, the majority of the current fake detection research still focuses largely on the English language. This research exists due to the abundance of available data (i.e., freely available large annotated datasets as well as already-trained English-language models). Nevertheless, the actual world information ecosystem exists in multiple languages. In many parts of the world, news content is produced and distributed simultaneously in multiple languages on the same platform or even in the same news posting, while low-resource languages have lower availability of data and are not well supported by current detection methods. As a result of their training implementation primarily occurring in high-resource linguistic environments, most of these models do not generalize well to multilingual/low-resource language environments.

An example of this issue can be seen with Urdu. With many speakers across South Asia, Urdu is heavily represented for communication in online environments across Pakistan and India. An example of this bilingual nature of much of the online content (news, social media) in those countries is that Urdu and English are used interchangeably to reach a wide range of people. Due to the bilingual nature of the content, determining the truthfulness of information using a fake news detection system can become very complicated as the fake news detection system needs to

maintain semantic consistency across both languages while correctly interpreting the original message. Additionally, there are not many large annotated datasets in Urdu available making it difficult for traditional text-based methods to accurately determine whether an article is false [3].

Most systems currently implemented have another limitation in that they focus primarily on unimodal data analysis. In real-world applications of fake news, it is uncommon that information being shared is presented solely in text format. Information being shared frequently is accompanied by images that provide additional persuasive characteristics. If all of a tweet's contents are evaluated and compared against other tweets, and no image is considered, then it would be easy to falsely conclude that there is a significant number of tweets in support of a given action. Images have a sizeable impact on emotions that people will feel based upon a particular piece of content (e.g., if someone sees an emotional image; they will likely feel a certain way about the content). Images can also provide a sense of reality to someone viewing the tweet; however, images can also be manipulated/modified and reused from other events, and/or can be put together with misleading text to misinterpret the content. Most models for detecting fake news on the internet ignore visual information, instead just focusing on textual indicators [4].

As a result, there is far less reliability in detecting fake news when only one type of indicator is used. Although a text might appear valid standing alone, the joint available photo could result in doubt concerning its validity. In contrast, an image by itself may seem to stand on its own but be repurposed to support the fictitious narrative created by the text.Detecting such inconsistencies requires an integrated analysis of textual information and images, as well as other media forms as appropriate.

In the first place, simply comparing common attributes shown by the features of either type of indicator will not suffice. In order to use both indicators to determine whether they support or contradict each other requires a more thorough understanding of what establishes a relationship between the indicator types as well as a thorough understanding of any finer areas in the two indicator types before the two indicator types can be used to aid in successful news detection.

In addition to needing to perform an integrated analysis of textual and images (as well as other forms of media that may be applicable), traditional multimodal fusion approaches generally employ a shallow integration approach (e.g. concatenation of either or both types of features together or late fusion) as opposed to a more complete merging and/or linking of actual features between the two types of indicators. While these approaches may marginally enhance performance, they do not allow for the deeper semantic relationship between the two modalities to be identified. In many instances of fake news, there isn't an obvious and direct contradiction between the two indicators, rather, there is a subtle misalignment, selective highlighting, or contextual manipulation. Therefore, applications that can identify such instances of fake news require more complex reasoning capabilities beyond simple correlations [5].

Humans have a natural ability to conduct these types of reasoning to judge the validity of the news they read. For example, when evaluating a particular news article, a person will instinctively contrast the words of the text against those in the corresponding photo to determine if the photo adds support or creates doubt as to the validity of the text within the newspaper. The human like cognitive processing involved in matching the two modalities includes determining if there is a discrepancy between them; attempting to find corroborating evidence from within both modalities; and establishing whether a particular piece of evidence is substantial enough to serve as justification across both modalities.

Because automated systems today lack the capability of conducting this kind of structured reasoning, there is often a disconnect between human evaluations of credibility and the predictions made by machines.

The above observations underscore the necessity of developing a new generation of multimodal, cognitively-based frameworks for detecting 'fake news'. Frameworks that can process information in a bilingual context, mitigate the challenges posed by low-resource languages and perform explicit reasoning about the relationships among text and images.

Furthermore, credibility assessment in multimodal fusion models should not only

be treated as strictly technical processes; detection models should be developed in such a way that the comparisons, verification and context-based reasoning method used by humans will be reflected.

This dissertation presents an effort to address these issues by creating a multimodal, cognitively-based framework called CogFusionNet. CogFusionNet combines transformer-based language models and vision transformers with cognitive-based reasoning methods to facilitate the melding of these modalities in order to provide a means of bridging the gap between human assessment of credibility and machine-based fake news detection. Ultimately, the aim of creating a system that has the capability of providing a secure, flexible, and explainable way to identify the presence of false information in real life; this includes assessing content from multiple languages/types.

1.2 The Impact of Transformers and Multimodal Fusion on the Detection of Fake News

The advancement of fake news detection systems has dramatically changed since the introduction of transformers, shown in Figure 1.1, (compared with prior sequential algorithms or fixed-size windows), transformers self-attention mechanisms enable models to better understand long-distance correlation and context relationships [6].This is significant for fake news, where misleading material requires an understanding of subtle context instead of simply recognizing false information.

When doing text analysis using transformer algorithms such as BERT, these algorithms have proven themselves to be very capable of representing linguistic subtlety, discourse structure, and semantic consistency within the linguistic domain of analysis [7].

As seen in Figure 1.2, by enabling the model to attend to all tokens in a sentence simultaneously, the ability to model long-distance relationships between word occurrences, capture latent meanings, and identify patterns associated with false/false/misleading information can be achieved. Due to this degree of context

sensitivity, they outperform traditional recurrent and convolutional network algorithms for complex classification problems where just using surface features will not provide sufficient data for effective classification.

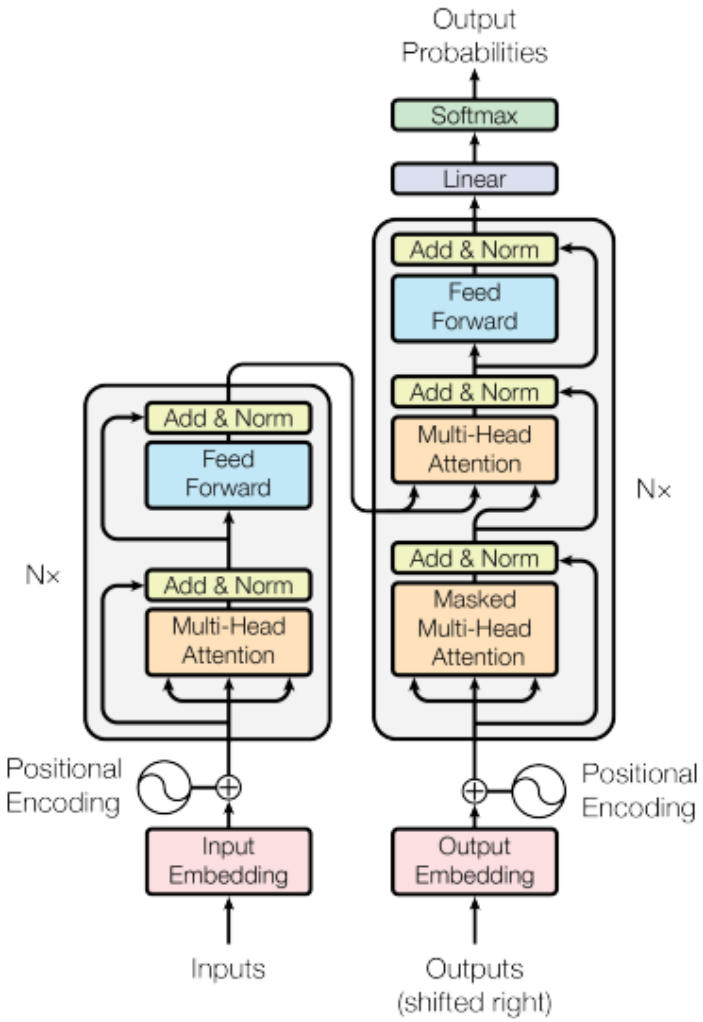


FIGURE 1.1: The Transformer - model architecture [6].

The role that transformer models play in multilingual and bilingual contexts is extremely relevant. Models like multilingual transformer models and language-specific adaptations (like UrduBERT [8]) allow for the creation of language representations that maintain semantic meaning but also take into account the differences in the various languages processed. Within bilingual contexts, transformer models are used to help close the semantic gap between the two languages so that detection systems can process content in multiple scripts and grammatical structures with similar consideration. This allows for an effective means to detect misinformation within multiple-language regions in the same information system.

In contrast to how text transformers are utilized to address the linguistic aspect of misinformation, visual information is equally vital to shaping perceptions of credibility. Images are also effective tools for persuasion by furthering a narrative, generating an emotional reaction, and providing a false sense of reliability.

Classic convolutional neural networks (CNN) have been used on a large scale for visual feature extraction but, due to their localized receptive fields, are not very good at capturing global relationships across images. For instance, in Figure 1.3, Vision Transformers (ViT) extend the capability of CNNs as they model the image as a sequence of patches and use a self-attention mechanism so that they can capture global contextual relationships across the entire image [9].

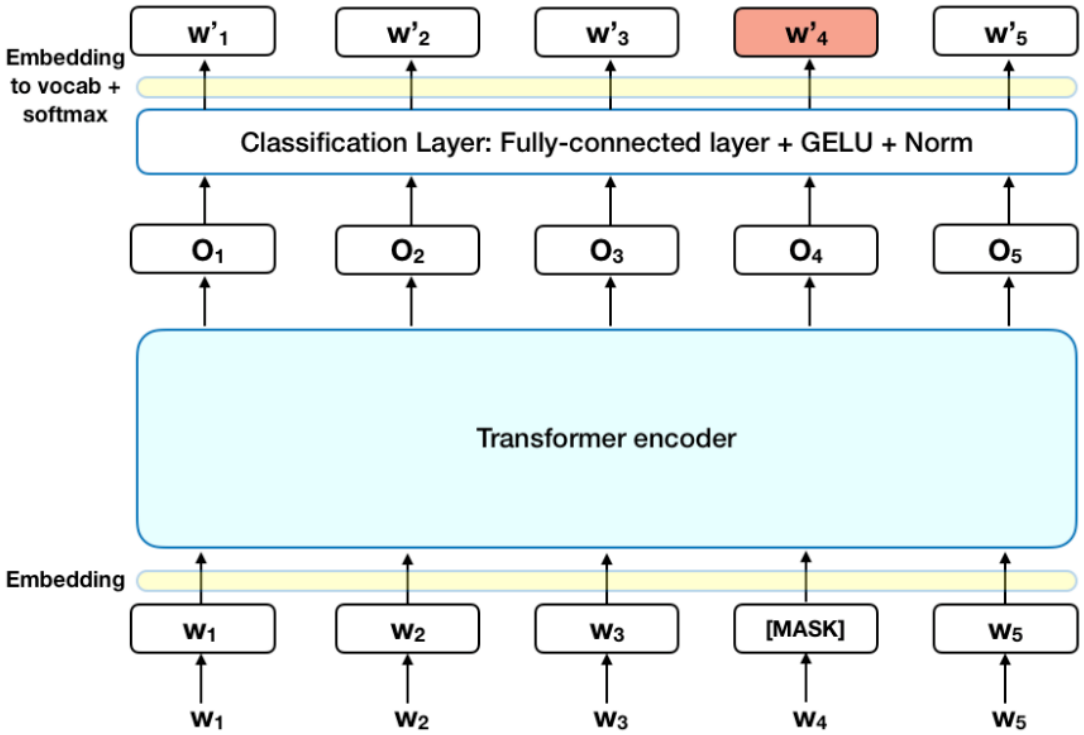


FIGURE 1.2: BERT Architecture

The use of visual transformers for detecting fake news allows us greater variability in visual representation, allowing us to better semantically compare those visual representations to accompanying textual content.

By capturing high-level visual concepts instead of low-level texture patterns, ViT-based models better align large high-level concepts of visuality with the textual

claim they are juxtaposed with. This is important because it allows for the identification of cases in which an image is reused, altered, or mismatched from its accompanying text.

Multimodal fusion is the underlying technology that integrates textual and visual information into one coherent representation of the news content. In the case of detecting fake news, it allows for the detection systems' models to become capable of not only analysing individual modalities but rather to develop a more complete understanding of the complete news story as it is told. With proper fusion, the detection system can establish whether an image supports the textual claim, contradicts the textual claim, or is irrelevant to the textual claim; therefore, establishing a relational understanding of information is critical to uncovering sophisticated misinformation techniques that are dependent on using one modality to delude users regarding the accuracy of another.

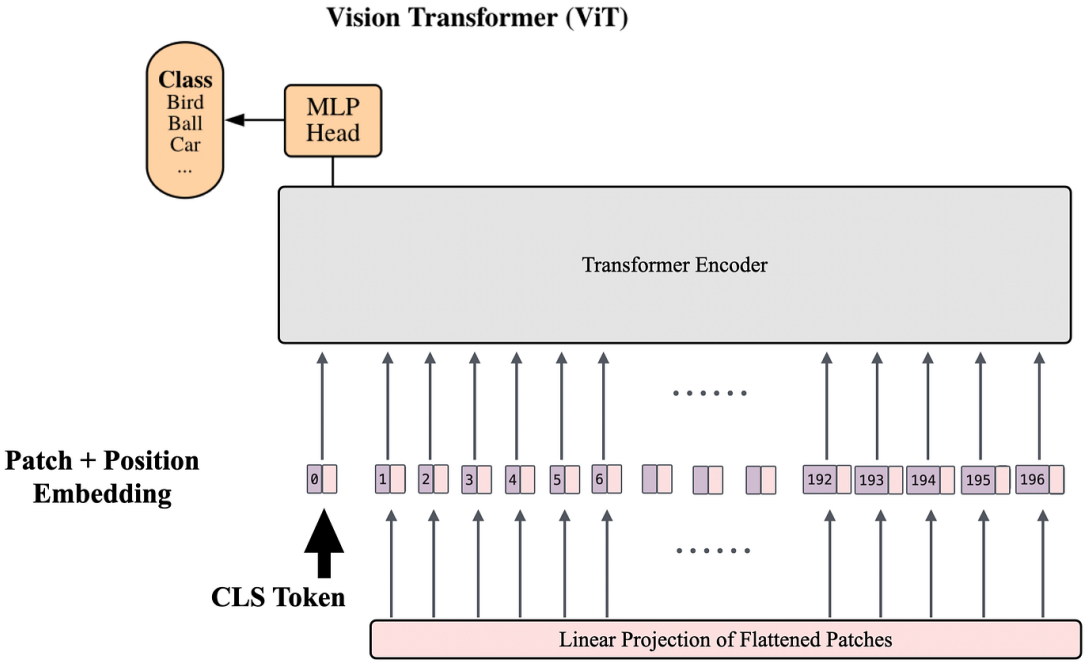


FIGURE 1.3: ViT Architecture [9]

Although multimodal fusion is necessary, it is not an easy task, as shown in Figure 1.4 and how text and images are derived from two different underlying feature spaces and how each encode information differently. As a result, simple fusion techniques such as concatenation of features or averaging of features often fail to capture true interactions between the two modalities. The methods mentioned

treat multimodal data as independent rather than interdependent sources of evidence. Therefore, they are limited in their ability to reason about the relationship between the different semantic types.

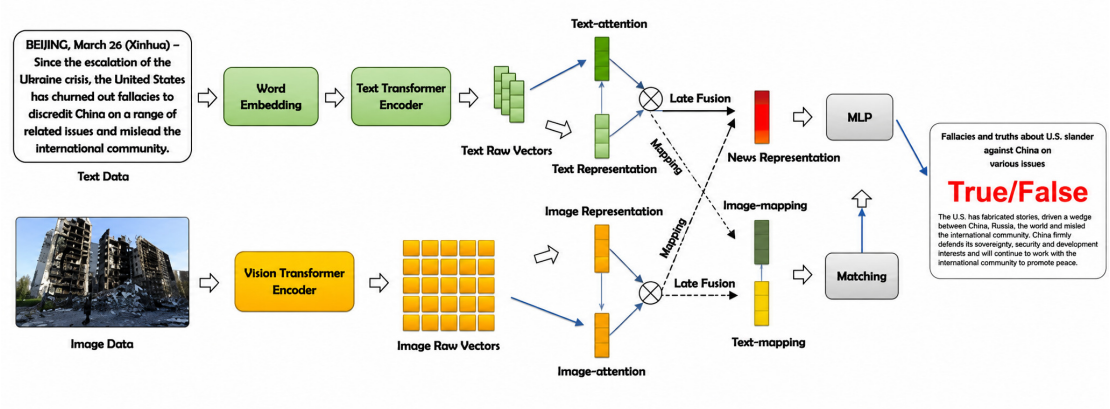


FIGURE 1.4: Multimodal Fusion Architecture

For that reason, shallow fusion techniques may fail to recognize subtle inconsistencies that would be immediately obvious to a human being.

The transformer-based fusion mechanisms, with their ability to perform cross-modal attention, can enable the model to learn which parts of the image relate to the specific textual token(s). Conversely, they can also enable the model to learn which parts of the text related most strongly to each specific image. By using attention to operate and selectively align across the modalities, the model can focus on semantically corresponding elements across the two modalities (e.g., matching textual references to visual objects/scenes). Thus, the use of cross-modal transformers provides an expressive and flexible framework in which to perform multimodal reasoning.

The role of multimodal fusion in fake news detection extends beyond enhancing task performance. Fusion mechanisms can also increase interpretability, as they can reveal how the different types of modalities contribute to the model's outcome decision. Attention maps and alignment scores can highlight how the model weighted the regions of the image and segments of the text when making the outcome decision. As a result, attention maps and alignment scores can provide insights into the model's reasoning process. This level of interpretability is essential for systems that work in sensitive areas, like misinformation detection, where

users require systems that are trustful and transparent [10].

The proposed CogFusionNet framework for transformer-based fusion was designed to mimic cognitive reasoning processes rather than exclusively relying on statistical correlations.

CogFusionNet replicates the human ability to evaluate credibility through comparing pieces of evidence across different modalities by means of specific structured pathways to combine language and image transformers. By allowing the system to think about when two different forms of evidence agree or disagree, the CogFusionNet is able to reason explicitly instead of learning it implicitly with only end to end training. In general, transformer architectures and multimodal fusing techniques will continue to be fundamental to advancing fake news detection toward more robust, adaptable & explainable systems. With the use of the capabilities of transformer architecture in both language and vision systems & using fusing techniques for both modalities, modern detection systems are able to address many of the issues associated with identifying real life misinformation. This thesis has taken the work described above and developed a cognitive & computationally based methodology which creates computational models of how humans make judgements about content in bilingual and multimodal environments.

1.3 Challenges Faced by Multilingual and Multimodal Fusion in the Fake News Detection

Challenges for multilingual & multimodal fusing are prevalent in the detection of fake news & while transformer-based models and multimodal learning methods have dramatically improved fake news detection, their ability to perform well in realistic multilingual or multimodal scenarios is still limited. The nature of how we see, hear and communicate allows us to interpret messages differently, even when using similar forms of media, like language and images.

The process by which we learn from that content is also impacted by our view of the way in which the content has been created, particularly when dealing with low-resource language (anything other than English written in an alphabet) as

there are very few annotated datasets (e.g., examples that have been previously classified) and pretrained language processing systems available in those languages. As such, these systems often struggle to achieve consistent performance across both languages and modalities (the format used to convey messages) [11].

The problem with the illustration of several concepts in one format results in difficulty for integration; there are also limitations in those systems. For instance, an individual engaging with such a system will not likely recognize a semantic contradiction among each of those representations derived from their use of visual representations generated through other contexts without the specific markings necessary to indicate their use outside of their intended context. Furthermore, when visually representing concepts across different cultures it is apparent that there is great variability between cultures' perception of visual representations and language creating an increased risk for misclassification. Additionally, transformer models require a very large number of computational resources to operate, therefore they cannot function in either "real-time" or "low resource" scenarios; thus rendering transformer systems unusable. Finally, there is no established procedure to formally demonstrate "explainability" for the multimodal systems currently being designed today, nor is there enough data to allow for establishment of such a procedure. Taken together, with new tactics for disinformation and the increasing disparity of available data, all of the aforementioned have the net effect of demonstrating low robustness and low generalizability across all multiple modalities. The need for frameworks that capture reasoning, consistency, and corroboration across languages and modalities comes from these challenges. This has resulted in an interest in cognitively inspired methods such as CogFusionNet, which aim to align the way machines detect fraud with the way humans identify credible information [11].

1.4 Cognitive-Like Mechanisms

Humans rarely consider only one source or modality of information when determining its credibility. They naturally compare text to associated images, determine if two sources of information either support each other or contradict on the same

people or events based on what they know about that person or event, contextually infer based on the information presented in a particular modality or medium, and rely on their prior experiences with similar types of claims to form an opinion about the credibility of new information. This thinking process of humans is not strictly linear, and it is not strictly follow-the-rules, but rather a combination of parallel thought processes, verification of evidence, and identification of inconsistencies.

As a result of the human thought process of a person using multi-source & multi-modal evidence when determining if something is credible; cognitively inspired mechanisms have emerged as a highly promising approach to improving the effectiveness of detecting fraudulent information, in particular, in multi-lingual and multi-modal environments. Most traditional machine learning (ML)-based approaches to detecting fraud focus on surface level features such as the frequency of words, the polarisation of the sentiment a word represents in a text, or visual patterns produced by the source of the content being analysed. While these methods can perform well on narrow datasets, they do not capture sufficient underlying thought process needed to determine the credibility of data in the real world as effectiveness will be reduced or completely ineffective given data used in the test build will not represent all real world scenarios. Humans do not just identify suspicious words or images; they assess the level of alignment between different modalities of expression and evaluate the logical consistency of the underlying evidence as well as the contextual plausibility of the information being presented. Mechanisms mimicking that of mankind's cognitive reasoning seek to reproduce these higher-level reasoning activities in a computational model.

At the heart of the cognitive reasoning approach to detecting fake news is the principle of integrating evidence. Mankind assesses various sources of information prior to drawing conclusions, evaluating both the relevance of each source and its reliability. In multimodal news posts, textual elements of a news post contain claims, narratives, arguments and so forth, while image elements provide either contextual context or emotional reinforcement for the same claims, narratives or arguments found in the text. A true cognitive-inspired system will treat

these different modalities as interrelated components of a single information entity as opposed to treating them as independent data sources. Thus, the cognitive-inspired approaches to detecting fake news will shift focus away from the mere concatenation of features found in text and images and toward the intentional use of cross-modal reasoning.

Another key cognitive processing function is inconsistency detection; when humans see conflicting evidence, such as an image that does not logically support associated textual information, they become suspicious about the validity of that evidence and its credibility. Detecting inconsistency based on visual and textual associations requires more than just statistical similarity; it requires semantic comprehension and contextual knowledge. Certain cognitive-inspired mechanisms model this reasoning process by detecting mismatches between textual assertions and visual identifiers.

As an illustration, consider an image that shows something unrelated to what the text claims regarding it. If you compare the meanings of the two separately, you may find they do not match up, even though superficially, the image seems authentic [12].

Beyond identifying inconsistencies, there also exist means to verify supporting evidence of corroborating that evidence. People tend to trust more when different forms of evidence can independently and mutually support one another. In a multimodal context, the establishing of credibility occurs when an image clarifies, supports, or strengthens a written account of what was observed. Cognitive-inspired models use a measure of the agreement between modalities to quantify how well aligned they are with one another, but they do not weigh all interactions in the same manner; rather, they give greater importance to reinforcements than they do to misleading or irrelevant signals [12].

In the CogFusionNet architecture, cognitive-inspired designs are based upon dual process reasoning through the creation of separate structured pathways to evaluate inconsistency and corroboration. These designs reflect those theories of cognition

psychology as they relate to when people determine whether something is credible; i.e., through both detection of conflict and establishing evidence to confirm the legitimacy of the information being evaluated. By incorporating these principles into the architecture of the models, CogFusionNet goes beyond existing attention-based multiplexing methods and takes a more interpretable, reasoning-based approach to multimodality fusion at the same time. The use of transformer models within a cognitive-inspired mechanism enhances the effectiveness of the cognitive-inspired mechanism as well.

Transformers are particularly suited for recognizing long-range dependencies and contextual linkages, both of which are necessary to perform cognitive operations on complex narratives and visual scenes. More than simply capturing local visual structures, Vision Transformers (ViT) can utilize important semantic regions of interest from images, thereby enabling the model to attend to semantically relevant parts of the image, rather than simply local visual patterns. When ViT is used in conjunction with transformers used as text encoders for multi-lingual content, this provides the ability for the system to share a common representation that enables cross-modal reasoning.

Similarly, cognitive-inspired architectures provide cognitive-based solutions for the multilingual challenges faced by the model in detecting fake news by focusing on the abstract nature of the information being analysed rather than relying on surface features or language specific patterns. Higher level concepts that are not limited by language, such as event and entity relatedness (engineering based approach), provide the model with an ability to reason across all language sites identified in multiple languages. This form of reasoning is therefore essential in the context of bilingual settings where there are both Urdu and English representations of the same content.

Interpretability is also another major benefit of a cognitive-inspired architecture, due to the cognitive functions driving the reasoning process, the model's decisions can be more clearly articulated and subsequently evaluated. By illustrating how different elements of the textual and visual content align or don't align, it helps users gain insight into why an article may or may not be deemed as 'fake' news.

Having this transparency helps to build user trust in auto classification algorithms, particularly in the sensitive settings of journalism or public policy.

In addition to offering a more transparent way to evaluate whether news is fake, cognitive-inspired solutions also give greater resilience against adversarial attacks on traditional classifications. Misinformation producers will often find it easy to create misleading text paired with an attractive photo that has no relevance to an article's text; thus, they may deceive systems that use only immediate fusions of the video/image content with the text content. However, systems relying on a more thorough evaluation of semantic relations and logical consistencies between video/image and text content will be more resistant to attacks based on purely visual similarities; therefore they provide for real cross-media relationships.

At a more general level, cognitive-inspired fake-news detection systems typify the move to 'human-centered' AI. They take into account not only the quality of outcome predictions but also the quality of reasoning, interpretability and contextuality of reasoning.

By aligning the cognitive strategies used by humans with the computational strategies used by cognitive-inspired fake-news detection systems, these same cognitive strategies are better able to account for: ambiguity; nuanced issues; and ever-evolving strategies of producing false/misleading news.

Cognitive-inspired solutions provide the theoretical foundation for improving the detection of multilingual, multimodal fake news. Rather than attempting to simply match the performance metrics of humans, the aim is to understand and integrate the strategies people naturally use when judging credibility.

CogFusionNet demonstrates that computational reasoning, when designed according to cognitive principles, can achieve both high performance and meaningful transparency, making it a viable solution for real-world deployment in linguistically diverse and resource-constrained environments.

1.5 Problem Statement

The growth of social media and online news has created a situation where misinformation can be disseminated very easily, which has revealed gaps in the current methods for detecting fake news. Even though there have been substantial advancements in this area, the majority of current detection systems have been developed to address English-language data and experience difficulties when utilized within multilingual and low-resource settings such as Urdu-English independent media ecosystems. This type of misinformation generally has a narrative written down that is paired with an image meant to persuade or confuse the target audience. Because of this, many of the detection systems that exist today are only focused on detecting false or misleading information through its text, as well as the few detection systems that are capable of incorporating images only do so in very simple ways and do not attempt to create connections between the two. Lastly, some models rely upon implicit attention without providing an explicit reasoning or interpretability, thereby creating models that do not indicate clearly whether or not visual and textual evidence support or contradict one another. The stated limitations show that there is a significant need for a better cognitive and trustworthy system with the ability to make rational judgments across different languages and different types of content. Therefore, this option proposes the development of a cognitive-based model that will accommodate both languages (Urdu, and English) for bilingual, fake news detection across multiple resources using many different types of tools (text & image).

1.6 Objectives of the Research

i. To create and implement a cognitive-inspired multimodal model that effectively fuses both textual (written) and visual (image-based) forms of information for bilingual fake news detection.

ii. To explore the role of advanced transformer-based encoders and explicit cross-modal fusion mechanisms in detecting inconsistencies and corroborations between text and images.

iii. To assess the durability and the ability of the proposed model to generalize across many different types of multilingual-multimodal datasets, using standard metrics for assessing performance and for determining how interpretable the output of the model is.

1.7 Research Questions

In order to streamline the proposed research, so that it can be systematically evaluated, this proposal will be conducted under several and very important research questions:

i. RQ1: How do cognitive-based fusion mechanisms improve the detection of fake news through explicit modelling of both consistency and inconsistencies between text and images in bilingual situations?

ii. RQ2: How well do transformer-based multilingual text encoders and visual transformers compare against traditional fusion methods for low-resource multimodal fake news detection?

iii. RQ3: How effective are the multimodal detection methods being developed at generalizing across languages and to what extent can the processes used to develop these methods be adapted to the types of real-world, multiscript disinformation messages (where text and image have only a loose relationship)?

1.8 Significance of This Study

The growth of the Internet and digital media has made the prevalence of fake news a major problem affecting many aspects of society, such as political context, orientation, and social coexistence. While there has been significant interest in developing automated methods for the detection and analysis of fake news in recent years, the vast majority of proliferation of methods for detecting and analyzing fake news is based on traditional text-only content in a single language (primarily

English), which does not provide much applicable assistance to those who interact with multilingual fake news content on a daily basis.

This study addresses this shortcoming by focusing on fake news detection in both bilingual and multimodal environments, and in particular, with respect to the Urdu-English fake news environment. Additionally, one of the unique contributions of this study is that it provides a strong focus on low-resource languages such as Urdu, which are typically omitted in computational datasets related to misinformation research. Urdu is spoken by millions of people in South Asia and around the world, and as such, continued challenges exist in acquiring and providing comprehensive knowledge regarding fake news in Urdu due to a lack of large, well-annotated datasets and well-validated detection methods. The novelty of this research is that it utilizes the Urdu Ax-to-Grind dataset in conjunction with a benchmark from English to help mitigate bias in language-based fake news detection literature. The second major contribution of this study is through its inclusion of multimodal misinformation studies that combine textual information with visual imagery to improve perceived credibility or distort perceived truth.

Many real-world instances of misleading images have been produced by reusing images or by slightly misaligning text and images rather than creating entirely new images that are purposely misleading. Detecting fake news with existing models is challenging since visual inconsistencies with text content are difficult to identify, especially when either data type is analysed in isolation or very little data fusion has occurred at the surface level. This research intends to establish a basis from which further investigation will occur into how visual evidence and textual evidence can be used in conjunction with one another to produce false information by providing the foundation for the development of multi-modal models that are able to simultaneously assess both visual and textual modalities.

This study is unique compared to many previous multi-modal studies because it applies an approach based on cognition, rather than relying on using attention as a way to reason through the process of determining whether to trust a source based on multiple forms of evidence (i.e., visual + text). This will allow for using verification and disproof of evidence from different types (or modalities) within

this study's design so that the study can take advantage of reasoning processes based on credibility, rather than simply evaluating correlations among surface features between modalities. As a result, the incorporation of processes involving reliability in reasoning into this design will lead to the development of systems for detecting fake news that are more robust and interpretable than current available models.

Additionally, having robust and interpretable systems for detecting fake news is critical for ensuring that fake news detection systems are implemented accurately in high-risk settings. From a methodological perspective, the study expands the current body of research regarding the ability to generalise language across borders by utilising real language aligned bilingual datasets to evaluate the performances of multilingual models using counterfeit directional translations or zero-shot models.

This design decision allows for a truer measure of multilingual robustness and allows for the discovery of bilingual and code-mixed specific challenges that arise within these language pairs which commonly affect the media landscape in South Asia. The results of this work are also important for policy makers, journalists, and platform moderators who need to be provided with reliable and understandable tools to help fight misinformation through the use of transparent reasoning methods and explainable cross-modal interactions. These reasoning methods and interactions will provide users with the necessary information to make intelligent decisions and thus increase their confidence in automated systems that they utilise. The results of this work will also be used to help create future datasets, models, and evaluate methods related to the attachment of multilingual/multimodal misinformation.

Overall, this study continues to support and develop systems that are more accurate, interpretable, and ultimately more socially responsible in their ability to identify instances of fake news. The advances made by combining bilingual text processing with multimodal fusion and reasoning expounded from cognitive theory form the foundation of fake news detection systems when applied to real world multilingual environments.

1.9 Scope and Limitations of the Study

This research project focuses on creating and evaluating a bilingual multimodal fake news detection system, using both textual information and images to assess credibility across two languages - Urdu and English - in South Asian digital media environments that use both of these languages individually and in parallel on a regular basis. By looking at both a high-resource language (i.e., English) and a low-resource language (i.e., Urdu), the goal of this project is to understand how robust fake news detection systems are across different languages as well as performance differences in multimodal fake news detection systems.

The study will utilize news articles that include both text and images because they represent a great deal of the misinformation found online through news websites and social media, where images are typically found alongside texts. The text will include headlines and content from the articles, while the images will come from either the lead image or an image that is contextually connected to the article content. The study's classification will be limited to a binary classification, meaning articles will be classified either as fake or real. This classification method is appropriate given how the articles are labeled in the benchmark datasets used for comparison.

The study will also research how well Transformer-based encoders can be used for textual representation and how well vision-based encoders can be used for image feature extraction in multimodal fake news detection. In this study, we propose a potential way of combining different sources of information (for example, the use of images together with text) by using an approach that models how the two pieces of information interact with each other to create a reliable representation of their combined value. By doing so, we intend to go beyond the traditional methods of combining features or attributes from both sources and instead provide an approach that models the attributes of each contributor and determines the trustworthiness of their combined output.

Our results are based on common classification metrics, which allow us to evaluate

and compare the results of our new method against previously existing state-of-the-art methods.

The datasets that we used in this study were derived from publicly available and published standard datasets (publicly available benchmarks), allowing for replicability and for fair comparison of our method to previous research. To approximate the real-world conditions under which image and text alignment do not always occur, we also introduced controlled noise in our datasets.

Although our study has contributed to the field, there are some limitations that deserve addressing. One limitation is the fact that our results are based on two languages, Urdu (an under-resourced language) and English; therefore, the applicability of these findings to other languages that are under-resourced or that have complicated morphology may not be entirely correct. While the findings from this research in a bilingual setting add important insights into developing a model that will apply to multiple languages, it will take additional efforts to collect new data from other languages, make linguistic adjustments, and validate the results.

A second limitation of our research is that we only used static images with news articles as the modality. Other modalities, such as video, audio, or user engagement signals (e.g., comments, shares, or propagation patterns), were not included in our analysis. The rapid rise in short video and digital content has made it impossible for this method to work effectively as there is insufficient time and no social context for people to rely on this type of evidence.

Additionally, each chosen dataset has been compiled and annotated by researchers who have deep expertise; however, because of how they are gathered, they may be subject to biases based off the domain/what the source of the dataset was/when it was collected etc.

That means that as the model is applied to domains where there have not been any datasets generated for a long time or to new topics that have come up in the news, or to present-day uses of false/misleading narratives, it may perform differently at those times.

As well, while this is true, the cognitive-inspired method used to combine the results of the model makes it more interpretable than other models of its type using standard attention-based models, it still represents a model approximation of human reasoning, as opposed to a causal explanation for the model's results.

In fact, while there are things that provide background support to the rationale used for making a credibility assessment, as well as things related to cultural context and intent, there are many things that cannot be modeled/computed precisely and fully.

Ultimately, the authors of this research are more concerned with developing a useful and effective model and a reasonable quality of reasoning regarding how credible something is than with developing a means to use the model efficiently (i.e., computing costs, amount of time for computing, and the amount of energy used in deploying the model are not accounted for, therefore future studies may examine developing lightweight implementations of the model to allow for this).

1.10 Contributions of This Research

This research is a contribution to the literature on detecting fake news in a multilingual, low-resource environment in that it solves many of the previous limitations with regard to these issues. One example is the development of a new bilingual multimodal framework called CogFusionNet for detecting fake news. CogFusionNet is built primarily for processing both English and Urdu written news articles as inputs for a fake news detector. This framework allows natively created bilingual data to be used rather than monolingual data or "synthetic" data that has been translated into either language; thus providing more effective and reliable evaluation of cross-lingual, fake news detection.

Another major contribution included in this research is the development of a bilingual multimodal benchmark, which was constructed by combining the ISOT Fake News Data Set (for English) with the Ax-to-Grind Data Set (for Urdu) in a manner that aligns textual news articles with photos contained in the same article. The resulting dataset allows for multimodal learning from both high and

low resource languages. Because existing multimodal fake news detection datasets are lacking, this benchmark represents an important advancement in this space, especially for Urdu.

Finally, this research proposes a new cognitive-inspired multimodal fusion technique,which is a departure from traditional cross-attention methodologies. By drawing on human credibility assessments, the fusion module proposed in this research explicitly distinguishes between how contradictory information is detected compared to how validation evidence is identified from text and from image. By doing so, it transitions from purely correlational learning of features to learning oriented towards assessment of credibility and better supports the assessment of deception or misleading information.

Furthermore, in addition to providing an improvement in terms of performance, the proposed solution also increases interactive satisfaction with the model by establishing discrete routes for processing each type of cognitive function via attention. This change provides much more visibility into the way that the model arrives at its conclusions, providing a greater sense of reliability and trust as well as suitability for a variety of real world applications, all of which require a credible level of performance in delivering transparent, trustworthy information.

This type of interactive satisfaction will be particularly valuable in fields such as public discourse, journalism and moderating activities for social media platforms.

Lastly, this study has validated the proposed framework on both English and Urdu datasets, thus demonstrating the ability of the framework to deal with the challenges of addressing linguistic diversity and data scarcity.

The results of the experimental work show the increase in accuracy achieved with cognitive-inspired multimodal fusion for the detection of misinformation, while still accounting for user satisfaction based on interpretability.

Collectively this research provides both practical and theoretical contributions towards creating credible, explainable and scalable systems for detecting misinformation across multiple languages and resource-poor environments.

1.11 Thesis Organization

The five chapters of this thesis are recognized by their diversity of content, where each chapter relates to a component of bilingual multimodal fake news detection and brings together a full examination of the study being examined. The organization will hopefully succeed at aiding the reader through the background of and reasons for conducting the research to methodology, results and conclusions obtained.

Chapter 1 addresses the problem of increasing difficulty presented by the amount of fake news currently being produced in the digital domain. Specifically, the study addresses the specific difficulty of detecting false news in a multilingual, low-resource setting. In addition, the first chapter describes the study's motivation, identifies the problems associated with the study, outlines the scope, limitations, key contributions and objectives of the research, the research questions addressed, and explains the importance of the research conducted. Consequently, the first chapter establishes the rationale and foundation for the research.

Chapter 2 contains a review of literatures related to the current methodologies employed for detecting fake news. This includes a review traditional approaches to text-based fake news identification, multilingual/cross-lingual based models, low-resource datasets, and multimodal identification models and approaches, as well as models based upon reasoning and explanations. The second chapter also identifies the existing gaps in the related literature which necessitated the development of the proposed cognitive inspired multimodal framework.

Chapter 3 presents the complete methodology for the research conducted for this thesis. It describes the datasets used in the research, how data was pre-processed, and the design of the proposed CogFusionNet framework (bilingual text encoders, visual feature extractor, and cognitive inspired fusion module).

Chapter 4 presents the experimental results pertaining to the proposed model's performance on either the English or Urdu dataset as well as comparisons of both

phases of experimental results to each baseline algorithm and state-of-the-art algorithms for either language. The chapter provides both quantitative results and qualitative analyses, in addition to a discussion of model behavior/robustness/interpretability.

Chapter 5 sums up all the findings and contributions of this research and contains a list of limitations related to this research along with future research recommendations for extending Cognitive-Inspired Reasoning/Multimodal Fake News Detection (MNFD) to a wider range of languages as well as to "real-world" applications.

Chapter 2

Literature Review

The manner in which individuals communicate with one another regarding the exchange and access of information has evolved significantly within the last 10 years. The introduction of Facebook, Twitter, LinkedIn, Instagram, Digital News sites and messaging apps, among other forms of social media, have enabled individuals to communicate much more quickly, access a broader audience than ever before and share information much more widely than was ever previously possible. Although this transition to digital communication has produced many advantages, it has also created a new environment in which false and misleading information can be disseminated relatively easily. Because of this digital change, fake news represents a serious and ongoing problem across the world.

Fake news most often refers to intentionally fabricated or distorted pieces of information that are presented in a manner consistent with authentic news articles. Fake news is much different than simply being inaccurate due to mistakes or misinterpretation as, in many cases, the goal of fake news is to mislead the reader, manipulate the reader's opinions and/or provoke the reader emotionally.

The effect of fake news can be reasonably large and will impact an individual's perception of their community, their community's government and the decisions made by government as well as their own behavior. Numerous examples exist in many venues, including election cycles, public health crises and social conflict, that have demonstrated how false narratives have increased confusion and negatively

affected the public's trust in generally believed and reliable sources of news and information.

Originally, combating fake news involved largely manual fact-checking by journalists and other independent groups. Such methods are effective at small scales but quickly prove ineffective against the volume of content produced online every day. Researchers therefore began looking for automated techniques that would allow for efficient and scalable identification of false information. As a result, fake news detection has become an area of active research in natural language processing and machine learning.

Early automating investigation methods focused primarily on analyzing the textual aspect of fake news. Those investigations were concerned with looking at the wording, use of language, writing style and statistical features of textual content in order to help distinguish between fake and true news. Researches were able to generate positive results with the automated processes performed in laboratory settings, however the automated processes were less successful when examining data from the real world.

The reason for this was because Fake news does not follow any fixed patterns, language can vary considerably depending upon the news topic, and target audience, as well as the cultural environment. This made it very difficult to generalize the research findings for purely text based media and the research was inconsistent, too.

As time went on, it became increasingly apparent to researchers that the text of fake news is rarely the entire story. Many fake news articles and social media posts contained images that were selected specifically to support the misleading claims. The images we see are generally unedited and come from various events that have been put together to create an image that may mislead us into thinking they come from the same event.

The inherent ability of visual media to assist in defining how someone sees an object, created a fundamental shortcoming in systems that did not use images as part of their overall analysis. This shortcoming led to a greater interest in the

development of multimodal fake news detection systems that assess both text and images together in their analysis.

Global social media has exposed another major limitation to conventional forms of false news identification — linguistic diversity. False news detection systems continue to have great success in providing accurate false news identification when measuring against English language databases; however; misinformation proliferates in lower-resource languages such as Urdu, where there is little to no access to openly labeled databases or linguistic datasets. Many users in lower-resource areas frequently write in multiple languages in the same post, and they often create posts that blend local languages with English in the same post or write a mixture of both languages in the same text.

Multimodality creates huge complication when the two are combined and measured, so multilinguality and multimodality create even more complications in the identification and classification of false news. Texts and images manifest semantics in fundamentally different manners, and the interrelationship between both is discerned through subtleties and indirect means.

Although the image associated with a news post may appear to relate on the surface level to the text, there are often underlying semantic contradictions. Detecting these inconsistencies requires more than just identifying patterns; it necessitates reasoning about context, intent, and the alignment of information across different modalities (i.e., the text and the image of the post).

As a result of recent research, there is a trend towards moving away from feature-based approaches to develop models that better approximate the way people use multiple sources of information to evaluate credibility. People often use multiple sources of information to assess credibility by comparing and looking for contradictions in the sources, as well as supporting visual evidence for their claims. Cognitive-inspired models aim to replicate these cognitive processes by explicitly modeling the interactions between text and images as opposed to treating them as separate, independent pieces of information.

Fake News Detection has evolved from simply classifying text to being a complex problem with many facets such as linguistic diversity, visual reasoning, and semantic alignment. Although researchers continue to work on this problem, most current models do not perform well in realistic situations, particularly for bilingual and lower resource languages. The difficulties in developing appropriate models highlight the necessity for more robust and interpretable methods that can jointly reason over both text and images similarly to the way humans do. Cognitively inspired multimodal frameworks provide a rationale for this research.

2.1 Text-Based and Traditional Fake News Detection

Two forms of detecting false news are traditional text-based methodologies or systems and traditional print-based systems. Historically, the detection of social media's false news has relied primarily on text-based detection processes, given that the primary source of news information for most online news articles is via text. Many early detection systems viewed the detection of false news as a text classification issue and therefore based their detection methodology primarily upon analyzing the characteristics of written language by employing linguistic and statistical parameters that exist within a piece of text.

The examinations of the use of n-gram features in conjunction with traditional machine learning classification techniques are perhaps one of the earliest detection systems created to study the detection of false news articles. The results from Ahmed et al. [13] indicate that false news articles can frequently be differentiated from true news articles through the lexical and syntactic characteristics of the articles themselves, with both lexical and syntactic characteristics measurable via surface-level text characteristics. This work helped to establish proof of concept for the potential of automated systems to accurately identify false from true news articles, with these automated systems using solely textual information as an indicator of whether an article is false news or true news.

Though the early text-based detection methodologies have shown success in distinguishing between false and true news articles, these traditional, text-based systems were hampered by their need to use manually-created features to classify news articles. The effectiveness of the n-gram-based detection methods was partially dependent upon both the quality and relevance of the features selected for use in the text classifiers, with these recalled features varying depending on the domain and topic the classifier was used in. Therefore, text-based classifiers based upon the use of n-gram features have been relatively successful in their detection abilities within specific contexts but have not performed well when exposed to articles written in a different context, timing, or instance, than those previously classified, nor have they had much success identifying the different ways that false news has evolved, is created, and then disseminated through the internet. Moreover, these models are insufficient to make a complete assessment of deeper semantic and long-term dependencies in text, which will help identify potential subtlety and context-driven instances of deception in text.

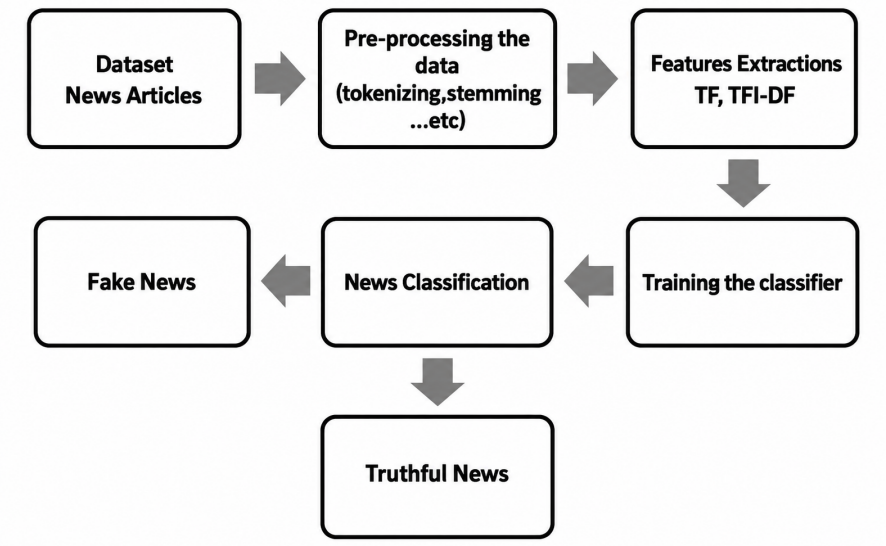


FIGURE 2.1: Early classification process of text [13]

Recognizing this limitation, later studies introduced the use of contextual and temporal information into text-based detection models. Agarwal [14] and colleagues utilized traditional textual analyses by implementing a spatio-temporal model of identifying fake news during the COVID-19 pandemic. Their study revealed that misinformation is not a static being; misinformation propagates and evolves through time as it continues to circulate. With their analysis of the propagation

of news content throughout social media networks and its temporal changes, their model demonstrated enhanced resilience for detecting misinformation, considering the very dynamic and crisis-laden environments. Their findings also demonstrated that supplementary contextual signals (i.e., temporal and propagation) provide value to enhancing the detection capabilities of textual features.

Despite the above temporal extension, text-only detection approaches are still insufficient for modern misinformation ecosystems. The majority of modern fake news also relies on visual components, such as pictures or video, to enhance the credibility and emotional impact of the story. Text-only detection methods do not account for visual manipulation, making them susceptible to multimodal forms of deception. In addition, many of these methods lack interpretability since the detection decision is based on statistical correlations, and therefore lacks explicit rationale. Consequently, while traditional techniques for detecting fake news were initially established based on text-only approaches [13, 14], they no longer are able to adequately address the exponential growth of complex, multilingual, multimodal, and real-world misinformation issues.

2.2 Low-Resource and Multilingual Datasets

There is an increasing prevalence of misinformation across many different language groups shows a need for additional resources to be used in the detection of fake news across those languages, particularly with regard to low-resource languages. English has a large supply of annotated data to use with robust pretrained models and ample NLP resources; however, languages like Urdu do not have such resources available to train on or develop generalized models. The lack of high-quality datasets severely limits how well the model can be trained and how well the model can perform in evaluating fake news detection systems in actual real-life multilingual environments [15]. Benchmark datasets for low resource languages have been created to help fill the void of these limitations. One of these benchmark datasets is the Ax-to-Grind Urdu dataset introduced by Harris et al. [15] which is the first resource established to assist researchers in the detection of fake news for Urdu. This dataset consists of news articles that have been annotated with

their corresponding metadata and whether they are true or false. This makes it easier to evaluate models built upon data from a language that has not received much research focus in the past. By using samples of real-world Urdu media, Ax-to-Grind demonstrates how complicated the patterns, reference points and styles of Urdu texts are. Having access to these datasets will also allow researchers to develop and refine their models for low-resource contexts, thereby bridging a big gap between high-resource and low-resource languages.

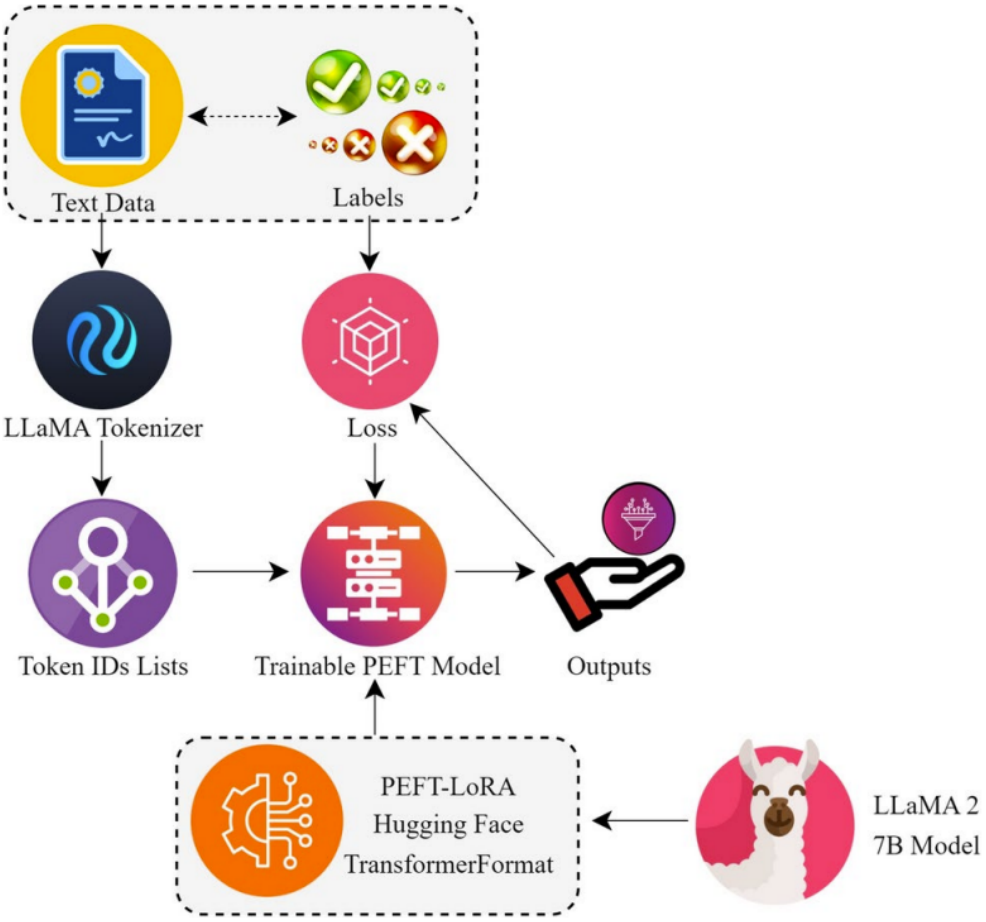


FIGURE 2.2: The proposed approach using PEFT [16]

In addition to this, the research team Harris et al. [16] created the **Hook and Bait** Urdu dataset, which expands on the concept of accommodating for different types of language and culture. Much like other "conventional datasets" that focus on a single subject area or category, the Hook and Bait dataset also includes a wide range of content across many subject areas in addition to including English-Urdu pairs. Thus, the Hook and Bait dataset offers an opportunity to test cross-language and domain-adaptive models with rigorously performed tests because of

its utilization of a diverse set of English and Urdu content and enables researchers to use large language models with cross-linguistic and domain-generating capabilities. Furthermore, Hook and Bait emphasizes real-world usability by employing content that includes naturally occurring noise and variation reflecting the real problems that model need to overcome.

Collectively, these datasets highlight the need for low-resource benchmarks to advance the field of fake news detection research. In addition to providing a baseline for the progress of developing models for use with underrepresented languages, they provide researchers with a common framework to use to compare their multilingual and cross-linguistic approaches to model evaluation. These datasets provide a basis to support the development of more powerful and generalizable systems similar to **CogFusionNet** using bilingual text encoders combined with multi-modal inputs for the problem of cross-language detection of false information in multi-language environments [15, 16].

2.3 Multilingual and Cross-Lingual Detection

While low-resource datasets have established benchmarks for measuring fake news detection, the problem of detecting fake news is greater than that of single-language models. Fake news is also commonly shared across languages in multi-language environments, which presents its difficulty for systems to be able to recognize, reason and generalize between languages. Cross-linguistic methods for detecting fake news address this problem by allowing models trained in one language to identify and classify fake news written in another language, which is particularly relevant for countries with large bilingual populations, such as in Pakistan and India, where there is a great deal of mixing of English and Urdu language news content [17].

BiL-FaND is an ensemble classifying approach proposed by Munir and Naeem (2019) to enhance the accuracy of identifying false information across both bilingual environments. BiL-FaND employs multiple classifiers that have been trained with datasets from both languages (a bilingual approach), and thus employs a

fusion strategy that improves BiL-FaND's accuracy over what would be possible by employing only one language. Using multiple classifiers allows the model to gather complementary information from both language classifiers, thereby decreasing misclassifications due to code-switching or partially translated text. The methodology supports the effective utilization of bilingual modeling in the development of low-resource environments and supports the use of ensemble strategies when developing detection systems to support robust detection of language features.

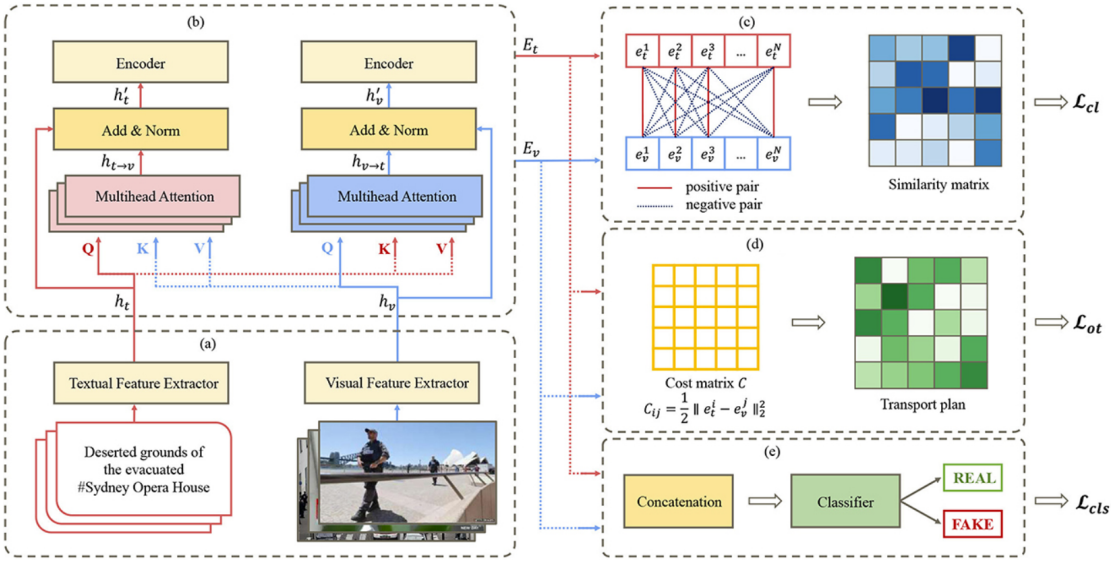


FIGURE 2.3: Multimodal approach [17]

Huertas-García et al. [18] developed a semantic-aware multilingual model that uses cross-language semantic representations to improve the flexibility of fake news detection systems. This framework also uses multilingual embeddings, maps semantic spaces across languages and allows models to learn from high-resource language data in low-resource language applications. This approach utilizes the significance of semantic consistency to develop cross-language detection models so that they can learn meaning rather than solely using surface-level activating features.

The cross-language transfer learning approach has been further developed by Han [19] who demonstrated that models trained in English can be retrained to support detection systems in low-resource languages (such as Urdu or Korean) via the use of fine-tuning and alignment of linguistic features.

This approach works well for low-resource target languages where training data is limited because the model can use extensive amounts of training data collected from English datasets and thus improve performance of the model in the target language.

Alghamdi et al. [20] present a hybrid summarization approach to low-resource language detection. A combination of content summarizing and multilingual Embeddings allow a method to capture both the preciousness of the content, as well as its cross-lingual correspondence, thereby enabling more accurate classification. This hybrid method allows for addressing difficulties with limited annotated data; it also offers additional interpretability for the model's predictions because of the summaries providing a compact rationale for the decisions made by the model.

Together, these studies point to the need for models capable of handling multiple languages, aligning semantic representations from language to language, and integrating additional signals through a strategy that is either ensemble or hybrid to be effective in detecting fake news in a multilingual environment. The developments outlined provide an improvement from generalisability over just textual data. However, the findings suggest that the visual aspect of the data has been neglected. This highlights the need for the CogFusionNet, which combines bilingual textual encoders with Vision Transformers (ViT) to improve visual reasoning. As a result, the CogFusionNet is being developed to form a more effective framework for multilingual multimodal fake news detection [17–20].

While there has been a lot of success with text-based methods of identifying fake news, real-world examples of misinformation are very rarely purely textual; rather, they will have images or videos or other visual material in conjunction with a misleading narrative to establish credibility for the content.

2.4 Multimodal Fake News Detection

While there has been a lot of success with text-based methods of identifying fake news, real-world examples of misinformation are very rarely purely textual; rather, they will have images or videos or other visual material in conjunction with a

misleading narrative to establish credibility for the content. Detecting fake news using multiple modes of data has become an important research field because of combining two different types of data (textual and visually based). This new approach also seeks to provide a framework for detecting fake news with greater accuracy by using different ways to build a model to learn how these separate forms of data correlate.

One example of this is KGAlign (La et al. [21]) which uses both semantic and structural types of knowledge to match up text and image data. KGAlign utilizes both CLIP (Contrastive Learning) and RoBERTa (transformer deep learning network) as well as the bottom-up approach of attention models: These models allow for the mapping of features from both text & image data into a common semantic space (defined as a graph). The result is that it provides the model with the knowledge necessary to draw conclusions about the relevant relationships between the text and image data when they are in congruence and whether there are any discrepancies between them.

Another researcher who has taken advantage of KGAlign is Xu [22] in developing a Multimodal Adaptive Graph-based Intelligent Classification Model (MAGIC). Xu's MAGIC model uses a graph-based approach to represent the dependencies of visually and textually based information (i.e., information that is clearly based on two different but related modalities). In this framework, each node is a semantic representation (e.g., a word or an image region) and each edge represents a meaningful relationship between two different modalities. Xu's framework has been demonstrated to enhance the ability to detect inconsistencies between text and images when multimodal interactions are explicitly represented at the structural level. Shen et al. [23] extended the existing multimodal encoding approaches using contrastive learning combined with optimal transport methods for enhancing the alignment of the two representations (image and text), while simultaneously reducing semantic differences between these two kinds of representation. This approach encourages the model to learn different representations of the same content so that the model can distinguish between a true correlation and the false correlations that result from cross-modal cues in challenging environments, where

misleading data are exploiting subtle features from both text and images.

Zhou et al. [24] proposed GS2F, which is a new method of fusion for semantically guided interaction between the two modalities (text and image). By combining cross-modal attention mechanisms and a structure for fusing multimodal data, GS2F allows the model to effectively assign different weights to relevant visual and textual features, thus ensuring that the identification of important inconsistencies is considered in the classification task.

In the context of low-resource multilingual settings, Bansal et al. [25] have created MMCFND, a system that integrates caption-aware multimodal features for Indic languages. This framework calls out the need to utilize any textual metadata (captions, alt-text, etc.) associated with images as a means of increasing the performance of detection systems in languages with limited annotation resources. By integrating textual features from multiple languages and visual features, they have shown that significant improvements can be achieved in both the accuracy of detection and the generalization across multiple languages.

Choi et al. [26] built CroMe, a unified system that utilizes cross-modal tri-transformers and metric learning to encode both text and image modalities. An architectural model specifically created to facilitate complex multi-modal interaction is CroMe, a scalable architecture that can handle the real-time flow of social media messages. This architecture illustrates how advanced attention mechanism may be employed to increase identification of subtle changes through the correlation between text and image thereby improving effectiveness for identifying subtle manipulation.

Finally, the authors Wei & Cao [27] proposed a similarity metric-based similarity-guided fusion for multimodal features' integration (image/text). The network uses this information about how close the visual content is, semantically speaking, to its associated textual content to help the network identify contradictory or misleading content — thus enhancing both interpretation and classification accuracy.

Together, these four studies demonstrate that multimodal feature concatenation through graph-based methods has evolved toward multimodal methods, which

include contrastive-based approaches, and ultimately transformer-based methods for detecting multimodal-based misinformation (i.e. fake news). Furthermore, they demonstrate the importance of integrating text and images in a manner that captures both the semantic alignment and the modality-specific nuances of both modalities. Despite these innovations, however, the challenges of cross-lingual applicability, real-time deployment, and interpretability remain; the CogFusionNet architecture addresses these challenges through its incorporation of bilingual text encoders, Vision Transformers (ViT), and cognitive-inspired cross-modal fusion mechanisms.

2.5 Explainable and Reasoning-Based Models

As the technology behind fake news detection advances, it has become increasingly difficult to not only detect but also to explain the reasoning behind why a model would make a certain decision about whether something is an instance of misinformation. Explainability is extremely important in high impact scenarios (e.g., political news, public health communication, and crisis situations) since automated decisions must be trusted and corroborated by human beings. Although multimodal transformers often yield high levels of performance for classifying instances of misinformation, the complexity of their architectures causes them to potentially operate much like black boxes and therefore limit interpretability and, more generally, diagnosing errors and biases.

To help improve interpretability and reasoning behind fake news detection under multimodal transformer models, Liu et al. [28] proposed the Modality Interactive Mixture-of-Experts (MIMoE-FND) framework. MIMoE-FND is based on the use of multiple expert networks with each expert specializing in one modality type and utilizing cross interactions between individual experts to incorporate the degree of cross-modal dependencies involved in the overall decision-making process.

Combining multiple expert assessments improves the reliability of identifying false content; from an evaluation of independent evaluation of modalities, a model that

combines the evaluation of each modality allows an evaluation of how each modality contributes to the overall assessment of the complete evaluation (consistent with the way individuals usually process information).

Ammal et al. [29] investigating and developing cognitive-based techniques using a multi-modal approach for evaluating false information that are very low resource languages and providing an example of how to create a specific consistency check based on large vocabulary word models and a patterned image of a given text. The model will identify both the location of inconsistencies as well as the reason for their occurrence and produce interpretable results to infer how the model reached its conclusion based on the input from a very low resource language text where there are no annotated data available to validate the conclusions based on a number of factors.

Explainability has been further enhanced by Xu et al. [30] through the incorporation of emotion-aware multimodal fusion in their AMPLE framework. Using emotional cues and attention-guided prompts, they indicate which parts of text and image contribute most significantly to predictions of misinformation. By overlaying attention maps with the contributions of specific modalities, AMPLE provides users with actionable explanations that may be useful in bridging the gap between model performance and user interpretation.

As a complement to these efforts, Athira et al. [31] conducted a systematic review of explainable AI techniques in fake news detection. In their study, the researchers identified various techniques (e.g., attention based mechanisms, saliency mapping) and architectures inspired by reasoning. In addition, the authors emphasized the need for transparency, fairness, and trust when using automated systems.

The results of the survey indicate that explainable models are useful for academic purposes but essential for deploying automated tools for decision-making by users and policy makers in real life. From the results of the studies reviewed, there is a consistent trend towards building multimodal methods of detecting fake (or misleading) news that use the combination of explainability and reasoning mechanisms as a fundamental approach.

By modelling the contribution of modality-specific resources, reasoning across modalities, and emotional or affective cues, these frameworks provide more than just predictive output, but offer explainable insights that can be validated, examined, and accepted as reliable by human users.

This body of research provides the conceptual foundation for constructing CogFusionNet, which is a cognitive framework that fuses vision and text inputs by using cognitive-inspired fusion and bilingual reasoning methods to provide high performance multimodal results with clearly defined processes and transparency [28–31].

2.6 Optimization and Efficiency-Oriented Models

As multimodal fake news detection algorithms continue to grow in complexity, concerns about scalability, computational efficiency, and practical deployment have become increasingly important. Transformer-based architectures such as those based on deep learning have high computational requirements, making them unsuitable for either real-time or large-scale use without developing optimization strategies. In an effort to minimize the overhead of producing false news detection output while providing high-quality output, researchers have focused much of their effort into creating frameworks that optimize and improve efficiency.

Ahmad et al. [32] and his team offered a model for the identification of false news that makes use of the particle swarm optimization (PSO) and reinforcement learning techniques combined to optimize transformer-based models. In this model, PSO provides a method for identifying the optimal parameters to utilize, while reinforcement learning is used to dynamically adjust the attention weights the model uses to focus on the most relevant multimodal features in the news article currently being processed. This combination allows for higher accuracy rates and also increases robustness due to an adaptive capability to dynamically prioritize the text and image features the model finds to be the most useful. The

result is reduced computation overhead by eliminating duplicate computations of multimodal features, thereby providing more efficient processing of large datasets.

Similarly, Li et al. [33] and his colleagues demonstrated the effectiveness of using Bayesian optimization methodology applied to CNN-BiGRU-based networks to generate predictions based on sequence data. Even though their research focused on predictions of sea level rise, the principles applied in optimizing hyperparameters through probabilistic search methods can be applied to detect false news articles. By using Bayesian optimization, we are able to allow our models to systematically explore the available hyperparameter space, thus obtaining the best performance with only a limited amount of training iterations. By using the above techniques on multimodal fake news detection, we can create accurate and resource-efficient models to help mitigate the impact of low-resource environments and devices with limited available processing capacity. When considering lightning-fast solutions to model parameter optimization, focus must also be put on efficiency through lightweight evaluation techniques and similarity benchmarks when working with edge-computing or real-time situations.

Liu et al. [34] proposed a technique for quickly determining the similarity of text documents in an edge computing environment by using a lightweight similarity method. Their similarity assessment resulted in accelerated assessments of semantic similarity by allowing for the use of compact representations, thereby greatly reducing the computational burden associated with using the compact representations, without significantly impacting the accuracy of their assessment. Such types of lightweight similarity assessments will be crucial to multimodal fake news detection systems that require simultaneous processing of both textual and image data and provide real-time decisions. Collectively the studies help to illustrate how critical it is to incorporate optimization strategies and hyperparameter tuning, along with lightweight computation into modern fake news detection systems. While sophisticated transformer-based and multimodal systems achieve high levels of accuracy, their feasibility lies in balancing between performance and efficiency. Optimization oriented methods represent an avenue toward deploying robust, cognitively inspired systems such as CogFusionNet in the real world to

support scalability, faster inference times, and the ability to adapt to different environments [32–34].

2.7 Datasets and Benchmarks for Low Resource Languages

High-quality datasets are a prerequisite to successfully Train, Evaluate, and Benchmark fake news detection models. Currently, there is a vast amount of data available (and thus, a plethora of annotated datasets) that researchers have investigated in relation to English-language corpora, but research has largely focused on finding datasets to use in low resource languages - for example, Urdu, Tamil, and various regional languages. Limited resources can hinder training on both models and the ability of rigorously evaluating their real-life performance across multiple linguistic contexts.

On the other hand, due to the historical lack of available resources such as annotated, domain-representative, and naturally occurring datasets, this has prevented the creation of strong fake news detection systems in these languages [35].

To overcome this issue, Shibu et al. [35] proposed a framework that empowers low-resource languages' fake news detection systems through large language models (LLMs). They emphasize that leveraging multilingual pretrained models will allow for the transfer of knowledge from high-resource languages to low-resource ones.

Furthermore, they explain that creating curated annotated datasets needs to be performed in order to collect the linguistic, cultural, and contextual dimensions of low-resource language misinformation. This will then allow researchers to evaluate their models on the conditions where they would be used, while also accounting for the different ways that misinformation exists in regional-specific contexts.

The study goes on to discuss how low-resource benchmarks are critical for both training and evaluating researchers' model building efforts. Researchers will be able to fine-tune their large language models to the specific low-resource language

using a structured dataset and evaluate the success of cross-lingual transfer, as well as evaluate multimodal architectures such as CogFusionNet.

The benchmarks demonstrated not only that the model can identify incorrect information in text but also generalized from one point to another regardless of different subjects, domains, or linguistic differences. In short, the creation and use of benchmarks for low-resource language datasets are necessary to develop a fair, solid, and scalable fake news-identifying system that can be applied in culturally different and multilingual real-life situations [35].

2.8 Conclusion and Research Gap

The current body of literature on fake news detection has made considerable progress in moving from text-only detection to developing detection systems that utilize multiple languages and modalities, as well as systems that incorporate reasoning.The traditional approaches to detecting fake news were primarily based on the textual features of the data using n-grams and various classical machine learning classifiers [13, 14] to classify fake news in instances where only controlled textual datasets were present.

However, the classical machine learning classifiers have limitations when it comes to generalizing the classification to other topics, domains, or for detecting misinformation in evolving tactics. Due to the increasing diversity of languages and types of visual media in use globally, research is now focused on low-resource and multilingual datasets [15, 16] and cross-lingual detection models [17–20] so that systems can adapt to different languages and domains. However, while substantial advances in the use of text for reasoning have been achieved, multimodal reasoning remains under-developed largely because text-only methods cannot adequately identify the more subtle semantic relationships between text and visual media.

Multimodal detection frameworks have recently emerged [21–27] and have begun to address some of these limitations by combining text and image data through the application of attention mechanisms, graph-based approaches and transformer architectures.

Table 2.1: Summary of Literature Review

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
1	Ahmed et al., 2017 [13]	Detecting Online Misinformation by n-gram-Based Machine Learning	Traditional text datasets	Using n-gram and ML classifiers (SVM, Naïve Bayes)	Differentiate real and fake news based only on the written patterns within the text	Only evaluates based on written text; unable to generalize across domains.
2	Agarwal et al., 2022 [14]	Spatio-temporal Methods to Detect Misinformation in COVID-19	COVID-19 related news	Determine how fake news spreads based on temporal features.	Real time captures share patterns of fake news over time.	Only evaluates temporal features; all other features are not evaluated.
3	Harris et al., 2023 [15]	Ax-to-Grind Urdu: Urdu-Language Fake News Dataset Uses	Urdu Articles	Revised Urdu-Language Corpus to detect Fake News	Provides an Annotated Low Resource Dataset for Detecting Fake Urdu News	No/ Limited capability for cross-lingual or multi-modal use.

Table 2.1: Continuation from Previous Page

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
4	Harris et al., 2025 [16]	Urdu News Hook and Bait Dataset	Uses both English and Urdu News	Domain-Agnostic and Multilingual Dataset	Improves cross-lingual evaluation & Supports fine-tuning large language models.	Requires the incorporation of multi-modal features.
5	Munir & Naeem, 2024 [17]	BiL-FaND: Ensemble of Bilingual Classifiers	Uses Urdu and English Language News	Ensemble of the Bilingual Classifiers	Improve Urdu/English Fake News Detection	Limited to text; no visual modality.
6	Huertas-García et al., 2021 [18]	Addressing Misinformation with Semantic-sensitive Multilingual Models	Multi-lingual datasets	Using Semantic Alignment; Cross-Language Model Transfer	Model Transfer to Make Knowledge Transferable Across Languages	Not Accommodating for Multiple Media, but Only for Text.
7	Han, 2022 [19]	Cross-lingual Transfer Learning for Fake News Detector	Low-resource languages & English	Transfer Learning for Fake News Detectors	Adapt in Low-Resource Language	Adapt to A Low-Resource Language Only in Text Format.

Table 2.1: Continuation from Previous Page

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
8	Alghamdi et al., 2024 [20]	Fake News Detection in Low-Resource Languages	Low-resource multi-lingual news	Multilingual embeddings + summarization	Greater accuracy and improved interpretation.	Limited multi-modal integration.
9	La et al., 2025 [21]	KGAlign: Joint Semantic-Structural Knowledge Encoding	Multi-modal news data (Images and Texts)	RoBERTa + CLIP + attention for text-image alignment	Enhanced semantic correspondence between images and text	Cross-lingual integration limited and focused on alignment.
10	Xu, 2024 [22]	A Multimodal Adaptive Graph-based Intelligent Classification Model	Multimodal datasets	Graphs to Show Cross-Mode Relationships within Data	Better and improved detection of subtle inconsistencies	Limited low-resource adaptation.
11	Shen et al., 2024 [23]	Multimodal Fake News Detection with Contrastive Learning	Using Multi-mode Data	Contrastive learning to align embeddings	Enhanced discriminative representations	Requires large-computational resources.

Table 2.1: Continuation from Previous Page

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
12	Zhou et al., 2025 [24]	GS2F: Multimodal Fake News Detection Utilizing Graph Structure	Using Multi-mode Dataset	Graph-structured fusion and cross-modal attention	Improved semantic interaction	Scalability for real time deployment unaddressed.
13	Bansal et al., 2024 [25]	MMCFND: Multimodal Multilingual Caption-aware Fake News Detection	Low resource Indic languages	Incorporating caption information into their models	Improved detection in low-resource languages	Still have limitations for cross-lingual generalizations.
14	Choi et al., 2025 [26]	CroMe: Multimodal Fake News Detection Using Cross-Modal Tri-Transformer	Multi-modal datasets	Tri-transformer networks to learn using metric learning	Handling of complex interactions between different modals	Potential issues with edge deployment remain unexplored.
15	Wei & Cao, 2024 [27]	Image-Text Similarity Guided Fusion Network	Multi-modal sentiment data	Similarity guided fusion network for sentiment-based sources	Enables identification of contradictory information	Did not focus on fake news, but on determining sentiment.

Table 2.1: Continuation from Previous Page

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
16	Liu et al., 2025 [28]	Modality Interactive Mixture-of-Experts	Multimodal datasets	Mixture-of-experts framework for cross-modal interactions	Modality-level and subject level interpretability	Resources for supporting low-resource languages are very limited.
17	Ammal et al., 2025 [29]	Reasoning Based Explainable Multimodal Fake News Detection	Low-resource Languages	Cognitive-inspired reasoning mechanisms and LLMs	Detecting inconsistencies across modalities	Datasets used had limited coverage and scalability.
18	Xu et al., 2025 [30]	AMPLE: Emotion-Aware Multimodal Fusion Prompt Learning	Multi-modal datasets	Emotion-aware multimodal fusion prompts& attention-guided	Enhanced reasoning and interpretability	Evaluation efforts on low-resource modalities were limited.
19	Athira et al., 2023 [31]	Systematic Survey on Explainable AI	Survey of XAI techniques	Categorizes various explainable AI techniques.	Highlights importance of trust and transparency	Just a survey, no new model proposed.

Table 2.1: Continuation from Previous Page

S.No	Authors & Year	Title	Dataset/ Platform	Contribution / Approach	Key Findings	Research Gap
20	Ahmad et al., 2025 [32]	Hybrid Optimization Driven Fake News Detection	Using multimodal datasets	PSO and reinforcement learning to optimize transformers	Increase robustness and efficiency	Still issues with high computational complexity.
21	Li et al., 2024 [33]	CNN-BiGRU with Bayesian Optimization	Sequence dataset	Bayesian hyperparameter tuning	Improve the accuracy of the model	Focuses only on sequence prediction, not multimodal fake news.
22	Liu et al., 2023 [34]	Lightweight Similarity Checking for English Literatures	Utilizing edge devices	Efficient and lightweight similarity computation	Enables real-time evaluation	Visual integration missing, text-based only.
23	Shibu et al., 2025 [35]	From Scarcity to Capability	Low-resource languages	LLMs & curated datasets	Enables structured evaluation & cross-lingual adaptation	Noted that multimodal integration is necessary for a holistic solution.

Additionally, these studies demonstrate that semantic alignment, cross-modal fusion, and caption-aware processing are particularly important in low-resource contexts [25, 26]. Nevertheless, there are still challenges to researchers in this area which are as follows:

i. Most models are heavily reliant on datasets derived from English (or other high-resource) languages, severely limiting their performance in low-resource settings.

ii. Cross-modal reasoning continues to be implicitly defined and thus many models do not explicitly identify evidence of inconsistencies or corroborations between text and images.

iii. Researchers commonly mention that a lack of explainable models is one of the main reasons why many people lose trust in the results generated by predictive analytics due to not being able to provide traditional meanings based on previous research to give them confidence that the predictions they receive from predictive analytics are correct and trustworthy [28–31].

iv. The areas essential for both real-time and large-scale implementations, such as multi-modal systems and transformer-based systems, are centres of interest with regard to computational efficiency and scalability [32–34].

v. Bilingual datasets of naturally aligned text-image pairs are limited, which reduces the ability of current models to operate in the ecologically valid environments [35].

The mentioned examples illustrate that there is one framework that effectively addresses the challenges posed by bilingualism, multi-modal fusion, explicit reasoning, interpretability and efficiency simultaneously. The ultimate goal of CogFusionNet is to utilize both bilingual textual encoders, such as BERT for English and UrduBERT for Urdu, and Visual Transformers (ViT) in order to address all these challenges.

In addition, CogFusionNet uses cognitive-inspired techniques for cross-modal fusion and explicitly models the task of detecting inconsistencies and verifying correlations across both modalities. Moreover, this architecture achieves alignment of semantic representation across multi-language modalities, provides interpretable reasoning pathways, while creating an improved sense of transparency and trustworthiness for the model's outputs. Lastly, optimization strategies applied to the model during training and inference phases facilitate its ability to be used in a real-world context in association with large-scale, multi-lingual social media content.

In conclusion, CogFusionNet directly addresses many of the critical gaps identified in the literature discussed in the preceding chapters in both an interpretable and scalable manner, thereby providing an alternative, robust framework for bilingual, multi-modal (i.e., visual and text) detection of fake news.

Chapter 3

Proposed Methodology

This chapter describes the methodological structure as shown in Figure 3.1 and 3.7 used for implementing the proposed design. Inspired by cognitive science, real world misinformation propagation scenarios are targeted where narrative built by textual and visual contents influence the audience misguidance. To improve the information credibility, Combine metalinguistic text representations with visual features through a fusion explicit mechanism known as CogFusionNet. This chapter discusses the methodology used to address these objectives, which includes (1) dataset selection and preparation, (2) modality-specific preprocessing pipelines, (3) model architecture design, (4) training strategies, and (5) evaluation protocols.

Two datasets that provide naturally aligned text-image pairs are used to provide ecological validity and avoid artifacts created from generating synthetic data. Transformer-based encoders are employed as text representations of the input, and vision transformers are used to extract image features; thus, both modalities have an identical architecture. To incorporate both similar and dissimilar relationships between the modalities, a cognitive fusion module is developed. Also, this chapter describes how experimental setups, performance metrics, and validation strategies were used to evaluate the performance of CogFusionNet, along with how they can be generalised to implement similar models. Altogether, these pieces come together to form a complete methodology for addressing the shortcomings of current

fake news detection systems that are multilingual and multimodal in nature.

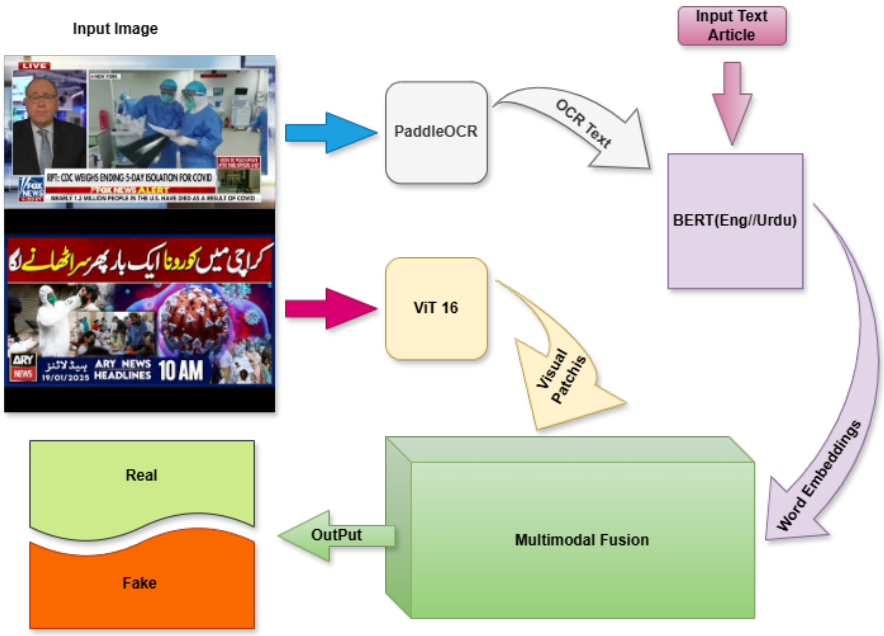


FIGURE 3.1: Block Diagram.

3.1 Dataset Description

The experimental evaluation of CogFusionNet is grounded in two publicly available and widely referenced datasets, selected to represent both high-resource and low-resource linguistic settings. Specifically, the ISOT Fake News Dataset is employed for English-language analysis, while the Ax-to-Grind Urdu Dataset is used to model misinformation in a low-resource context. To create an equal and manageable basis on which to build an experimental corpus of English articles taken from the ISOT Fake News Dataset, we extracted a sample of 10% of the total dataset in order to create a balanced (2:2) sample. The presence of this clear separation of true and fake news articles allows simple labelling of the articles while also retaining the diversity and realism associated with real-world news.

3.1.1 English Dataset (ISOT)

The news articles used for developing this corpus are contained in two separate files, each representing a different class of news (true and fake). The presence of this clear separation of true and fake news articles allows simple labelling of the

articles while also retaining the diversity and realism associated with real-world news.

3.1.1.1 Dataset Composition

There are 44,919 news articles. Each has records associated with it and includes metadata or textual information compiled in two files as follows:

i. `Fake.csv`: Contains 23,502 news articles labeled as fake

ii. `True.csv`: Contains 21,417 news articles labeled as genuine

Thus allowing samples within each class of news to include a full breadth of features for feature extraction.

3.1.1.2 Data Attributes

Each news article within the two files has the following attributes:

i. Title: The title of the news article. The title of the news article contains the headline of the news article. The title of a news article is commonly written in such a way that it contains much of the semantic and emotional content of the news article, and it will also attract readership. As such, the title is a significant source of information for the identification of fake news articles through sensationalised/misleading framing.

ii. Text: Within the text column, we provide the complete content (i.e., full body of news article). The text provides the linguistic data for the article, and provides substantial amounts of contextual information to conduct in-depth semantic analysis. The amount of text and representation of it greatly varies from article to article, mirroring the diversity of real-world reporting mechanisms.

iii. Subject: This attribute represents the thematic category of an article (e.g., politics, world news, government, or social issues). The subject allows us to identify topical patterns related to misinformation and enables the model to learn domain-based characteristics of fake/real news.

iv. Date: The date attribute identifies when a particular article was published. Temporal references are valuable in identifying trends within news reports and have relevance to news reports linked to specific events or periods where misinformation increases.

3.1.1.3 Label Assignment

Class labels are implicitly defined by the source file:

i. Articles from `Fake.csv` are assigned the label *fake*

ii. Articles from `True.csv` are assigned the label *real*

This labeling strategy ensures consistency and avoids ambiguity during training and evaluation.

title title of news article	text body text of news article	subject subject of news article	date publish date of news article
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	 2016-01-13 2017-12-31
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. m...	politicsNews	December 29, 2017

FIGURE 3.2: ISOT Dataset

To create a corresponding visual element for each article in the subset, articles were processed by scraping lead images from original article URLs (when available) and alternative images embedded within articles (when lead images were not present). Data collection of natively aligned text-image pairs allows the researcher to maintain the natural relationship between textual claims and attached visuals and to avoid creating artificial and/or unrelated text-image pairs. The resulting dataset consists of natively aligned text-image samples and closely reflects how news content appears on actual online platforms.

3.1.1.4 Relevance to Fake News Detection

The dataset is useful for research related to detecting fake news because it's both large and diverse. This dataset's strength derives from it providing both text and additional context to support the study of fake news detection strategies. The dataset does this by being collected in the same way as typical news and has real examples of information that is misleading as it relates to the study and use of fake news in this field of research. There are both titles and complete articles provided in the dataset so there are various options for researchers to test how different types of analysis work. For example, researchers may want to test headlines only, body text only or do a combination of both body text and headline analysis.

The subjects of articles in the dataset cover a wide range of topics, including, but not limited to, the following; political issues, public health issues, technology and science, entertainment and social issues. The articles cover a broad range of topics and provide a broad range of different writing styles, visual frameworking, and narrative purposes. Therefore, the variety of articles in the dataset will add to the complexity of the work of fake news detection and provide a way for researchers to evaluate the generalization ability of specific multimodal models, especially because misinformation will typically be different from writing styles, visuals and narrative purposes.

In conclusion, this dataset is a very strong dataset to train and evaluate machine learning and deep learning-based models for classifying fake news, and can be used

to develop traditional text-based methods and the newer transformer model-based methods.

3.1.2 Urdu Dataset (Ax-to-Grind)

In terms of its overall impact on society, the rise of false information through digital channels or platforms has severely affected public perception, confidence in the institutions that service them and the stability of political systems at both national and international levels, particularly in areas of the world where the rate or volume of information consumed from the internet exceeds the speed or rate of fact verification mechanisms. There are several fact checking projects and databases available for English language information; however, access to high quality resources (i.e., datasets) written in less widely spoken languages such as Urdu is still very limited. As Urdu is one of the most commonly spoken languages throughout South Asia, it has been largely overlooked in terms of academic (research) endeavours in the area of fake news detection due to the relative lack of available high volume quality annotated datasets.

```
Sr. No.,News Items,Label
1,ٹی ٹی پی نے پنجاب حکومت کے پبلی کاپٹر کے عملے کو یرغمال بنانے کی تردید کی ہے۔,FAKE
2,مارک زکربرگ سیاست میں آنے کا سوچ رہے ہیں۔,FAKE
3,فریدہ جلال نے اپنی موت کی افواہوں پر تنقید کی۔,FAKE
4,جعلی خبریں: پاپ اسٹار حدیقہ کیانی نے جعلی منشیات کے بارے میں برطانوی ویب سائٹ پر تنقید کی۔,FAKE
5,صنم ماروی نے میڈیا پر گردش کرنے والی زیادتی اور ٹکیتی کی کوشش کی افواہوں کی تردید کی۔,FAKE
6,نواز شریف نے واقعی کرپشن کی وجہ سے ڈیووس میں تقریر کرنے سے منع کیا۔,FAKE
7,بی ایف اے کی 'آئرن لیڈی' عائشہ ممتاز کا ویڈیو- کریشن الزامات,FAKE
8, CPEC کے بارے میں خطرناک افواہیں,FAKE
9,امریکہ نے القاعدہ کی جعلی ویڈیوز بنانے پر کروڑوں خرچ کیے جسے بیورو آف انویسٹی گیٹو جرنلزم نے بے نقاب کیا ہے۔,FAKE
10,پاکستان کا جعلی پروگرام نشر کرنے پر بھارتی چینل کے خلاف قانونی کارروائی پر غور,FAKE
11,بھارتی صحافی نے تصدیق کر دی کہ فوج نے کوئی ثبوت فراہم نہیں کیا۔,FAKE
12,گلگت بلتستان میں بھارتی میڈیا کی جعلی بغاوت بے نقاب,FAKE
13,نواز شریف واقعی دنیا کے دوسرے کرپٹ ترین سیاستدان ہیں۔,FAKE
14,مسئلہ کشمیر پر وزیراعظم نواز شریف کے آج قوم سے خطاب کرنے کی خبریں بے بنیاد ہیں۔,FAKE
15,یونیسکو واقعی اسلام کو امن کا سرٹیفکیٹ دیتا ہے۔,FAKE
16,ان کا کہنا تھا کہ میری جوٹ جعلی تھی: حفیظ نے ٹی وی چینلز کے خلاف شکایت درج کرادی,FAKE
17,پاکستان کے باتول بھٹو کا خطرناک جھوٹ,FAKE
18,ہم جنس شادی کے قانون سے متعلق جھوٹی خبروں کا پردہ فاش,FAKE
19,نوازدین صدیقی کا کہنا ہے کہ جھوٹی خبریں پھیلانا بند کریں۔,FAKE
20,چنئی میں ٹیسلا کے پہلے ہندوستانی یونٹ کے قیام پر وائرل جعلی خبریں,FAKE
21,اے آئی سی سی کے ترجمان کا کہنا ہے کہ بی جے پی جعلی خبروں کی سب سے بڑی فیکٹری ہے۔,FAKE
22,"لاک ڈاؤن کی جعلی خبریں وائرل، زلزلہ آگیا",FAKE
23,ٹاٹا موٹرز نے ایم جی دومکیت ای وی کا مذاق نہیں اڑایا,FAKE
24,مراٹھی اداکارہ جوئی گنگری نے اپنے حادثے کی افواہوں کی تردید کردی,FAKE
25,تیلگو کامیڈین سدھاکر نے اپنی موت کی افواہوں کی وضاحت کی۔,FAKE
26,دھیریندر شاستری پر جعلی خبروں کے لیے یوٹیوب پر مقدمہ درج,FAKE
27,دہلی ہائی کورٹ نے یوٹیوب چینلز کو جعلی خبریں شیئر کرنے سے روک دیا,FAKE
28,نانک محل دارووال حویلی کوئی مذہبی مقام نہیں ہے۔,FAKE
29,کیا لندن پولیس نے عمران خان کے بیٹے کو گرفتار کر لیا؟ سوشل میڈیا کی گپ شپ ختم ہوگئی,FAKE
30,جین نے 800 سال پرانی مسجد کو تباہ کر دیا۔,FAKE
31,پنجاب حکومت نے بشریٰ بی بی کے داماد کو ٹیوٹا کا چیئرمین مقرر کر دیا۔,FAKE
32,میں زیادتی یا قتل کیا گیا تھا۔ GCWUF طالب علم کو,FAKE
33,بھارتی میڈیا بالاکوٹ فضائی حملے میں '200 سے زیادہ ہلاکتیں' ثابت کرنے کے لیے پاکستانی فوجی کے جنازے کی ویڈیو استعمال کرتا ہے,FAKE
34,یالٹ شہز الدین بھارتی فضائیہ کے ساتھ جھڑپ میں جان کی بازی ہار گئے۔ بھارتی پراپیگنڈہ جھوٹی خبر نکلی۔ F-16 پاکستانی,FAKE
35,سوشل میڈیا پر اپنی موت کی جھوٹی خبریں پھیلانے پر ایک شخص نے صحافی کی پٹائی کر دی۔,FAKE
```

FIGURE 3.3: Ax-to-Grind Urdu Dataset

To fill this void and address the specific need for a quality dataset capable of being

used as a benchmark in the field of fake news detection in the Urdu language, a new dataset known as the Ax-to-Grind Urdu dataset has been generated and made publicly accessible. This resource represents one of the most extensive efforts in the creation of a corpus that supports research related to misinformation in Urdu language contexts. Additionally, the Ax-to-Grind Urdu dataset serves as a foundational dataset/resource necessary for the development and evaluation of machine learning/deep learning models in low resource settings.

TABLE 3.1: Dataset Distribution with respect to Numeric Features.

Features	**True**	**Fake**	**Combined**
Unique Words	12,894	25,176	29,911
Average Words per News Item	34.82	116.98	75.90
Average Characters per News Item	171.79	528.62	350.58

With 10,083 news items, Ax-to-Grind Urdu is the largest publicly available multi-domain and cross-domain dataset for Urdu fake news detection. The dataset contains a total of fifteen distinct domains, such as politics, health, sports, entertainment, technology, weather, agriculture, economy, showbiz, social media, education, women's rights, religion, foreign affairs, and international news. This wide variety of topics allows for models trained on the dataset to encounter multiple differences in linguistic styles, narrative structures and types of misinformation that are specific to their domain.

The news articles were collected from the websites of many of Pakistan and India's largest newspapers and news networks over a six-year period (from 2017 - 2023) the two countries with the greatest populations of native Urdu speakers. The geographic diversity of the dataset helps provide a more realistic dataset and decreases regional bias, allowing for cross-domain and cross-context analytics. All news articles contain naturally occurring content and not synthetic or translated content; these criteria are critical in evaluating real-world systems for detecting fake news.

The dataset's annotation was conducted by professional journalists; thus, a high level of reliability and factual accuracy was assured. Each news article is assigned either a fine-grained label of either "True" or "Fake," depending on verification from reliable online sources and fact-checking methods.

The research community widely considers the Ax-to-Grind Urdu dataset as an important dataset for developing accurate machine learning systems to detect fake news in Urdu. It was introduced in detail in a study accepted for the 22nd IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-2023) which will help to enable the systematic evaluation of fake news detection models in Urdu and to promote additional research into multilingual, cross-lingual, and low-resource misinformation detection. The Ax-to-Grind Urdu dataset provides authentic, expert-annotated Urdu news content and is essential in providing a sound basis for testing bilingual and multimodal fake news detection frameworks. The large scale, diverse content, and high-quality annotations provided by the Ax-to-Grind Urdu dataset make it well suited as a resource for assessing the linguistic nuances, cultural context, and domain-specific misinformation patterns often overlooked by researchers working with English-centric datasets.

3.1.3 Dataset Quality and Noise Analysis

A manual review of 500 randomly chosen sample sets from the overall combined data set will allow an assessment of how well the text and image will align together. Approximately 82 percent of these pairs contained meaning or content as shown through the image; meaning that the image provides direct support, visual aids or context for the message contained in the text. The remaining 18 percent were poorly aligned mode of information such as re-used visual modes of information from other types of events and generic images that don't clearly relate to the textual message. The sets of improperly aligned images will not be discarded from the data set because in real life often times misinformed visual data includes non-related visual data or noise. By including cases like this it will allow the model to learn how to be robust against imperfect multimodal alignments. Also,

this inclusion of noise will help to minimize the tendency to fit perfectly idealized conditions. The data will be used to provide a more realistic and challenging benchmark for the comparison of CogFusionNet.

3.2 Text Preprocessing and Linguistic Normalization

Effective textual preprocessing is a key factor when working with transformer-based models, especially when those models must also be able to support multilingual and under-resourced participants. There are many types of raw linguistic inputs that contain components of both noise and inconsistency. When these raw linguistic inputs are taken from online news articles or social media posts, they often include noise, inconsistencies and artifacts related to the respective platforms that will cause negative consequences for the model performance if not addressed appropriately.

To avoid these potential detrimental effects on model performance, we applied a systematic approach to the preprocessing of English and Urdu textual inputs, separately, prior to passing the content through the language encoders utilized in the CogFusionNet model.

3.2.1 English Text Preprocessing

Preprocessing of the dataset for the English language focused on removing non-content elements (i.e., anything introduced during web scraping) from the articles, while still providing clean, consistent information in the article text. Articles were stripped of HTML tags, embedded scripts and advertisements, navigation text and any redundant whitespace, inconsistent line breaks or formatting symbols to ensure consistency in the text's structure. The next step was to convert the entire article text into lowercase letters for the purpose of reducing vocabulary sparsity and allowing the model to treat semantically identical words consistently. Special characters or excessive punctuation were only retained where they helped

define the article meaning (e.g., quotation marks, question marks) or possibly indicate rhetorical framing or sensationalism (i.e., common features in misleading material). Stopwords were not intentionally excluded. In a traditional machine learning approach, many stop words could be removed, but this is not the case for a transformer-based model (BERT), which benefits from complete sentence structure since the function words are very important for syntactic and contextual understanding. Stemming and lemmatization were also not performed because pre-trained transformer tokenisers utilise subword tokenisation to manage morphological variance.

Once cleaned, the English text was tokenized using the BERT tokenizer, which employs WordPiece-based subword segmentation. This allows the model to effectively represent rare words, named entities, and previously unseen tokens—an important property when dealing with evolving misinformation narratives.

3.2.2 Urdu Text Preprocessing

There are several additional challenges in preprocessing Urdu text. The first is the script characteristics of Urdu language. The second is the morphological richness of Urdu and the last is the lack of standardization of Urdu text on the Internet from many different online sources. The fact that Urdu is written in the Perso-Arabic script makes it difficult to maintain consistency in terms of spelling, diacritics, and spacing of words. Because of this, a preprocessing strategy that was designed specifically for the Urdu language was used. To preprocess Urdu text, first, all non-Urdu characters were filtered out. Non-Urdu characters were defined as any characters that were not in the Urdu script including Latin letters and symbols and emojis that did not provide any semantic value. In contrast, code-mixed content (Urdu and English are used interchangeably) was preserved and not removed because it is commonly used in actual Urdu language news and social media posts. If these were removed, the ecological validity of the corpus would have been significantly reduced. The second set of normalizations included common orthographic inconsistencies, such as different forms of the same character occurring and irregular spacing between words. Since diacritics (marks used in

writing to indicate different sounds) in Urdu are not consistently used in online text, they were also removed as a method of limiting the sources of variability that are produced without changing the meaning of the text. Stopword removal as well as aggressive morphological processing was avoided as well throughout the corpus so as to maintain sentence level coherence. Urdu-language tokens were created by using the UrduBERT tokenizer. The UrduBERT tokenizer was pretrained to accommodate the morphology and structure of Urdu script. Because it segments text into meaningful subword units, it allowed the creation of robust representations of a low density of vocabulary items.

3.2.3 Sequence Length Management and Padding

In order to standardise input dimensions across batches, we set a maximum sequence length for both languages; sequences of text that exceeded this length were cut off, such that the first parts of these text inputs would remain intact since headlines and first paragraphs are generally where the most significant claims are made in news articles. Because of this decision, shorter length sequences were padded with special tokens in order to have equal-sized inputs across all of the examples for both languages. Attention masks were also produced in conjunction with tokenised sequences to indicate which tokens represented actual characters and which were padding. The purpose of the attention masks was to allow the transformer encoders to focus computational effort on tokens containing real meaning, while ignoring those with padding fold during self-attention computations.

3.2.4 Rationale for Minimal Text Manipulation

A major design consideration for the preprocessing pipeline was to use as few aggressive text manipulations as possible; rather than completely redesigning textual features using advanced techniques, the design strategy was based on how transformer-based encoders perform best (by learning rich semantic representations of text via the use of raw or only lightly modified data), which also reduces the amount of information lost and supports improved generalizability, particularly with respect to subtle language cues, rhetorical techniques and implicit claims

that are often made within fake news stories. By consistently applying language-aware preprocessing steps, any textual input to CogFusionNet will retain high levels of semantic fidelity while being able to be used with both English and Urdu transformer models. This foundation is essential for effective cross-lingual comparison and for reliable fusion with visual representations in subsequent stages of the framework.

3.3 CogFusionNet Architecture

The overall architecture of CogFusionNet, a cognitive-inspired bilingual multimodal fake news detection framework, is illustrated in Figure 1. Our proposed model aims to detect misinformation in both English and Urdu by using the respective languages' news articles as a source of information, as well as images and associated text from those articles for analysis. To achieve this goal, we will not treat the two modalities (text and images) independently; rather, we will design CogFusionNet to model the relationship between both modalities as it relates to how people typically consume and interpret misinformation in the real world. To do so, we will use a modular architecture—for example, each component will focus on a single modality yet provide data for a combined decision-making process. At a high level, CogFusionNet consists of three primary components: (1) a dual-language text encoding module that encodes text into two languages (English and Urdu) using language-specific transformer models, (2) a visual feature encoding and optical character recognition (OCR) feature extraction module, and (3) a multimodal fusion mechanism based on the cognitive principles of how humans make judgments of credibility. The text encoder encodes the news articles' text in each language, and the visual encoder extracts the semantic representations and OCR representations from the images that accompany the text. The encoded representations of the text and images will then be fused together using a fusion module based on the cognitive principles of human judgment of credibility (i.e., comparing visual evidence to assertions in text to identify contradictions and verifying corroborating cues) and then assessed through the classification head of CogFusionNet. Essentially, CogFusionNet uses a structured way for text and image features to

interact prior to producing the final output.

3.3.1 Text Encoder

In CogFusionNet, a model for understanding how textual information can be used across different languages to maintain linguistic nuance and semantics, news articles are decomposed into their two parts: headline and body text. Headline and body text are then jointly analyzed to derive meaning based on context. Because English and Urdu fundamentally differ grammatically, morphologically, and lexically, the model uses separate pretrained representations for both languages.

For processing an article written in English, a pretrained BERT-based encoder is employed because BERT has demonstrated superior performance with respect to holding long-distance dependencies and providing context-based representations of word relationships within English-speaking contexts. Conversely, an article written in Urdu utilizes an UrduBERT (a variant of the transformer architecture) pretrained specifically for use within low-resourced language environments on behalf of Urdu corpora. This choice of transformers offers the opportunity for the model to learn patterns specific to individual languages such as inflectional morphology, script form, and semantic indicators that may not be captured using traditional multilingual encoder models.

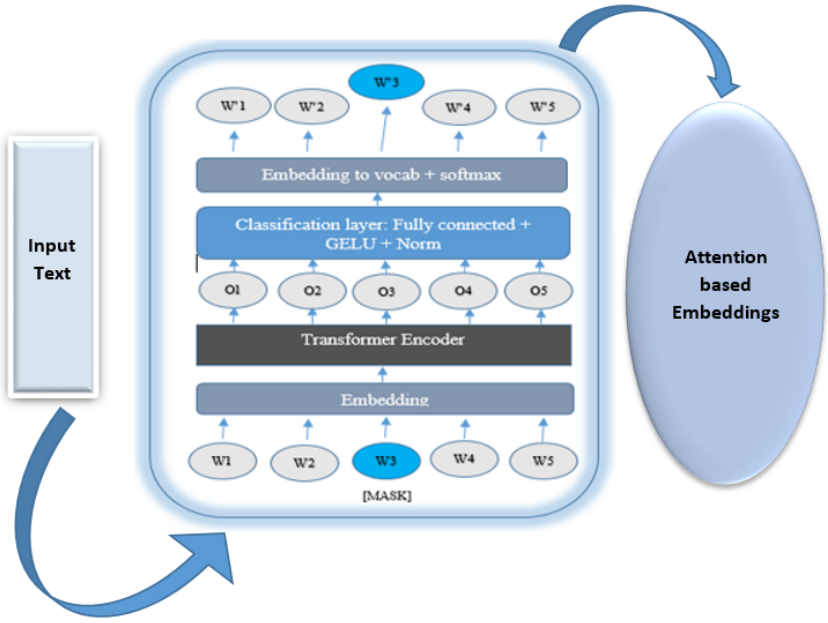


FIGURE 3.4: BERT architecture.

Both transformer languages produce token-level embeddings that are contextualized thereby reflecting each token's meaning in accordance with its surrounding tokens rather than only with respect to itself. Therefore, the model is able to interpret nuance as it relates to negation, emphasis, and implicit claims which are commonly found within evidence patterns of misleading narratives. To obtain a fixed-length representation for each article, token embeddings are aggregated using a combination of CLS-token projection and mean pooling across the sequence.

3.3.2 Image Encoder

In CogFusionNet, image processing associated with news articles is performed through a Vision Transformer (ViT) backbone (ViT-Base/16), which provides evidence of good performance for capturing long-range dependencies and contextual decoupling in visual data. In contrast, ViT metadata situates various convolutional networks as they have segmented images (and these segments form the tokens) into individual 16x16 (stated) and there are no overlapping sections to them. In order to represent the images in the transformer, the segmented images are linearly projected onto embedding locations and positional encodings are added to position each on the transformer where the image has been segmented.

Let the input news image be represented as:

$$I \in R^{H \times W \times C} \tag{3.1}$$

where:

i. H, W are image height and width

ii. C is the number of channels (typically C=3)

The image is divided into non-overlapping patches of size $P \times P$ where:

$$P = 16$$

The total number of patches is:

$$N = \frac{H \times W}{P^2} \tag{3.2}$$

Each patch is flattened into a vector:

$$x_i \in \mathbb{R}^{cP^2}, \quad i = 1, 2, \ldots, N \tag{3.3}$$

The ViT transformer's encoder will perform multi-headed self-attention on all image patches to model all of the interactions that have occurred within the image, providing the opportunity for all image patches to identify regions of high interest and extract object-level characteristics that may be indicative of whether or not a news article is credible. For instance, if a news article has an image that has been edited in a way that indicates manipulation, it could mean there may be inconsistencies in the object arrangement of the image, scenarios that are unlikely, and edits that do not appear to be consistent with the text related to the news article. By capturing both global context and granular visual information through the image representations, the ViT architecture will provide a rich, higher dimensional representation that the cognitive fusion module may utilize to inform cognitive fusion by bringing about both implicit and explicit inconsistencies and corroborations between the visual and text sources.

Each flattened patch is linearly projected into a D-dimensional embedding space:

$$z_i = x_i E \tag{3.4}$$

where:

i. $E \in \mathbb{R}^{CP^2 \times D}$ is the learnable projection matrix

ii. D is the embedding dimension (ViT-Base: $D = 768$)

To preserve spatial structure, positional embeddings are added:

$$Z_i^{(0)} = z_i + p_i \tag{3.5}$$

where:

i. $p_i \in \mathbb{R}^D$ is the positional embedding for patch i.

A learnable CLS token is prepended:

$$Z_0^{(0)} = x_{CLS} + p_0 \tag{3.6}$$

Final input sequence:

$$Z^{(0)} = [z_0^{(0)}, z_1^{(0)}, \ldots, z_N^{(0)}] \tag{3.7}$$

For each transformer layer $l = 1, \ldots, L$ the sequence is updated as:

Multi-Head Self-Attention:

$$MSA(Z^{(l-1)}) = Concat(head_1, \ldots, head_h)W^O \tag{3.8}$$

Each attention head is computed as:

$$head_j = softmax\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right) V_j \tag{3.9}$$

where:

$$Q_j = Z^{(l-1)}W_j^Q, K_j = Z^{(l-1)}W_j^k, V_j = Z^{(l-1)}W_j^V \tag{3.10}$$

To create a comprehensive fusion of multiple modes, the output of the ViT encoder (a feature vector produced from the CLS token) can be transformed into a different latent space for compatibility with the BERT and UrduBERT generated textual embeddings, which will allow the cognitive fusion module to compare across modalities meaningfully and combine them. A visual representation of the

information encoded semantically as well as information that may provide potential misleading evidence or supporting evidence in relation to bilingual creation of fake news.

Residual Connection + Layer Normalization:

$$\hat{Z}^{(l)} = LN(Z^{(l-1)} + MSA(Z^{(l-1)})) \tag{3.11}$$

Feed-Forward Network (FFN):

$$FFN(z) = \sigma(zW_1 + b_1)W_2 + b_2 \tag{3.12}$$

$$Z^{(l)} = LN(\hat{Z}^{(l)} + FFN(\hat{Z}^{(l)})) \tag{3.13}$$

Global Visual Representation (CLS Token):

After L transformer layers, the **CLS token** encodes the global image representation:

$$V_{CLS} = Z_0^{(L)} \tag{3.14}$$

This vector captures:

i. Global context

ii. Object-level interactions

iii. Visual inconsistencies or manipulations

Latent Space Projection for Multimodal Fusion:

To align with BERT and UrduBERT embeddings, the CLS token is projected into a shared latent space:

$$V = W_v V_{CLS} + b_v \tag{3.15}$$

where:

i. $W_v \in R^{D_f \times D}$

ii. D_f is the multimodal fusion dimension

Final Output of ViT Backbone:

$$V \in \mathbb{R}^{D_f} \tag{3.16}$$

This visual embedding derived from the previous step is passed to the Cognitive Fusion Module, where it is evaluated against the text embeddings received from BERT and UrduBERT to determine if the news content provided by the two modalities is legitimate.

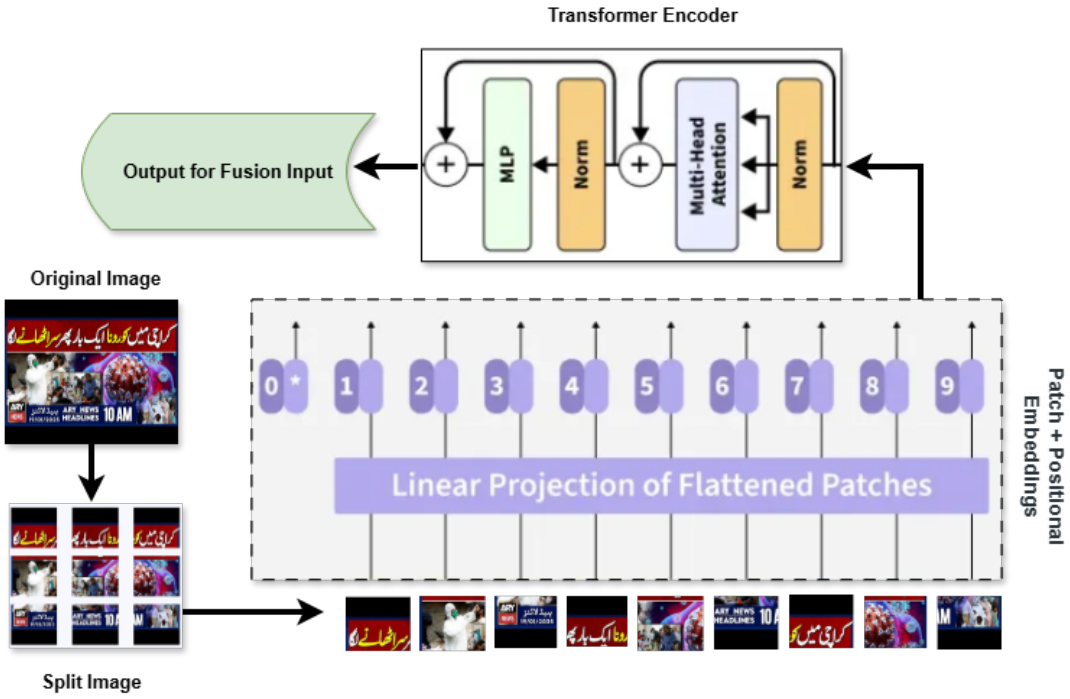


FIGURE 3.5: ViT architecture.

3.4 Textual Information Extraction Using PaddleOCR

Using images that accompany news articles in multimodal fake news detection, there are many times when the image contains text information, such as a headline, caption, slogan, banner, date/time, or numerical claim. This image-based text

often has a large impact on how readers interpret the news content and is often used as a means of misleading readers through false information in fake news. In order to directly capture this text-based information, CogFusionNet includes PaddleOCR, which is specifically designed as a text extraction module that will provide image text in news articles.

PaddleOCR, unlike traditional image-based feature extraction modules that predominantly focus on extracting features associated with objects in images, has the ability to recover the language-based textual content within an image. This is particularly important in fake news scenarios, where images may appear visually plausible but contain misleading or fabricated textual overlays that contradict or falsely support the accompanying article text.

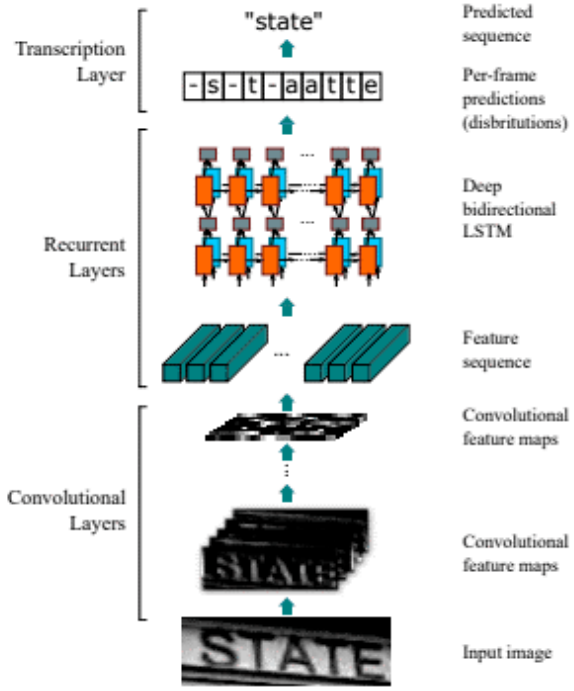


FIGURE 3.6: PaddleOCR architecture.

3.4.1 PaddleOCR Pipeline

PaddleOCR is designed with three main stages allowing for the processing of images and their contents in three different ways. These stages are Text Detection, Text Recognition and Text Direction Classification (optional). This design allows for strong overall performance and ability to have the same quality across many different types of images when processing news articles across the web.

3.4.1.1 Text Detection Stage

In the first stage the PaddleOCR Text Detection Module identifies the location in which there is a high probability of text content in an image. This detection module can successfully detect text when it comes in many different sizes, orientations and fonts, which is frequently the case for real-world news images like posters, screenshots and social media graphics.

PaddleOCR uses a segmentation-based detection approach to create candidate text regions by separating text pixels from background pixels, then generating bounding boxes that surround individual words or lines in the candidate text regions. By only passing the regions around potential text areas on to the next module for further analysis, this detection module greatly reduces noise from non-text visual elements.

3.4.1.2 Text Recognition Stage

Once the text regions have been identified, the individual text regions will be cropped and given to the Text Recognition Module. The purpose of this module is to convert the visual representation of the text region into a sequence of readable characters. PaddleOCR uses a sequence-based method to recognize text that is very strong against a wide range of font styles, scales and levels of complexity in the background.

The recognition process produces textual strings corresponding to the content embedded in the image. These strings may include event names, slogans, names of public figures, dates, or numerical values, all of which can be highly indicative of a news article's credibility.

3.4.1.3 OCR Output Aggregation

All of the recognized text segments from all of the different sections of the image are combined as one text segment. This combined text segment of the image is

the same visual text story that is derived from the image and can also be used to complement both the story text and the visual object-level features found in the image as indicated by the Vision Transformer. By preserving the original ordering and grouping of all the recognized text segments, the system is able to preserve the context of all of the recognized text segments so that a person examining the story can make a meaningful comparison with the textual claims made by the story.

3.4.1.4 Integration with CogFusionNet

The extracted OCR text is treated as a type of additional text mode. The OCR tokens that are designated as being extracted by the OCR engine will be processed similarly as the article text is using the same language specific encoder used on the article text to maintain a similar semantic representation as well as allowing for the OCR derived content to interact smoothly with both the text-based and visual-based features within the same modality.

The OCR information is particularly valuable in cases where:

i. The image contains captions or headlines that reinforce or contradict the article text.

ii. Visual text introduces new claims not explicitly mentioned in the article body.

iii. Misleading images rely on fabricated textual overlays to appear credible.

3.4.2 Role in Cognitive-Inspired Fusion

In the cognitive-inspired fusion module, OCR-derived text plays a dual role:

i. Inconsistency Detection: OCR text is compared with article content and visual features to identify contradictions, such as mismatched dates, locations, events, or entities.

ii. Corroboration Verification: When OCR text aligns semantically with both the article text and the image content, it provides supportive evidence that strengthens the credibility assessment.

Additionally, it creates a similar type of behavior to a human fact checker, because a person usually looks at the text that is found within the image in order to verify or question if the news story is valid.

3.4.3 Benefits of PaddleOCR in Fake News Detection

The inclusion of PaddleOCR enhances CogFusionNet in several ways:

i. Enables extraction of hidden textual cues from images

ii. Improves detection of cross-modal inconsistencies

iii. Strengthens robustness against deceptive visual manipulation

iv. Supports multilingual fake news analysis, aligning with bilingual research objectives

Collectively, PaddleOCR enhances the multimodal cognitive process of CogFusionNet by connecting visual and textual understanding, leading to high levels of accuracy and interpretability in detecting fake news.

3.5 Cognitive-Inspired Fusion Module

CogFusionNet's cognitive-inspired fusion technique reflects human evaluation of news credibility. Traditional multimodal fusion techniques rely on concatenating text and images or on a "one-size-fits-all," generic type of cross-attention. However, this approach is fundamentally different in that it simulates human processes of reasoning. The fusion method for integrating, comparing, and verifying textual and visual input data has been decomposed into two unique attention pathways based on the two cognitive functions of contradiction detection and corroboration verification.

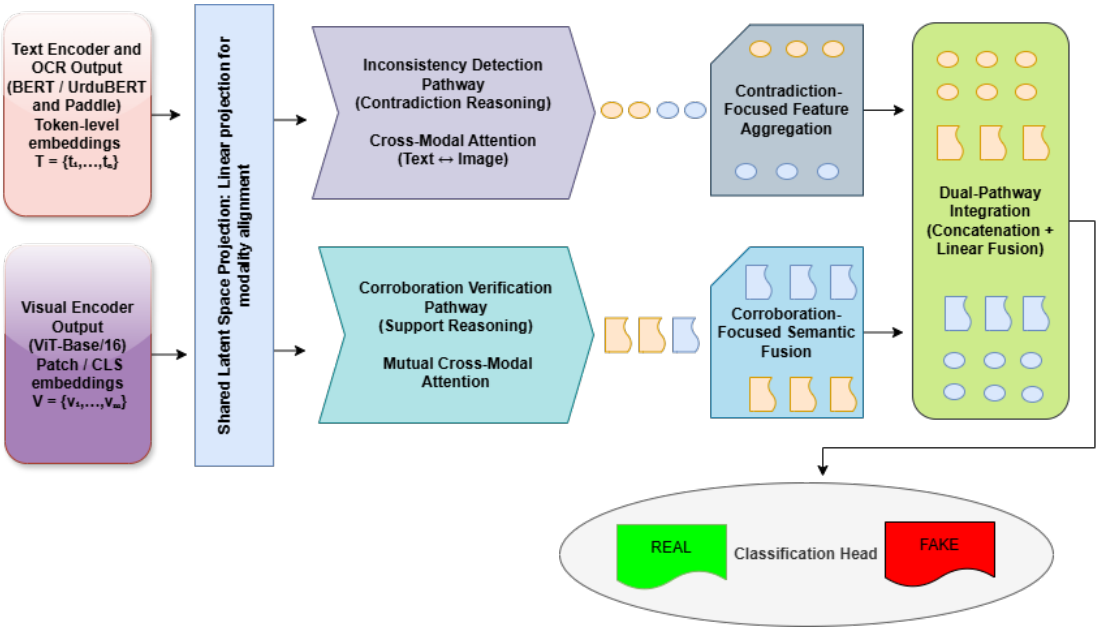


FIGURE 3.7: Multimodal fusion (proposed).

3.5.1 Input Preparation

The fusion module accepts text representations (embeddings) produced by the transformer model specific to each language (e.g., English: BERT; Urdu: UrduBERT) and visual (image) descriptions using the Vision Transformer encoder. Text embeddings for all 5 languages comprise sequences of contextualized word features, while visual embeddings describe both object details and scene details from the associated image.

Both text and visual embeddings are projected into a common latent space using sequential linear transformations prior to merging with the fusion module for the purpose of providing meaningful interaction between modalities.

Let the textual embeddings produced by BERT or UrduBERT be

$$T = \{t_i\}_{i=1}^{N}, \quad t_i \in \mathbb{R}^{D_t} \tag{3.17}$$

and the visual embeddings extracted by the Vision Transformer be:

$$V = \{v_j\}_{j=1}^{M}, \quad v_j \in \mathbb{R}^{D_v} \tag{3.18}$$

Both modalities are projected into a shared latent space of dimension d:

$$\hat{t}_i = W_t t_i, \quad \hat{v}_j = W_v v_j \tag{3.19}$$

3.5.2 Inconsistency Detection Pathway

The first pathway, the inconsistency detection pathway, is specifically designed to identify contradictions between text and image. It operates as follows:

i. Cross-Modal Attention: Each text token attends to all image patches, and vice versa, using a multi-headed attention mechanism. This enables the model to detect where textual claims and visual cues do not align semantically.

$$A_{inc} = softmax\left(\frac{\hat{T}\hat{V}^T}{\sqrt{d}}\right), \quad f_{inc} = Pooling(A_{inc}) \tag{3.20}$$

ii. Semantic Alignment Scoring: The attention outputs are combined with a learned similarity scoring mechanism to highlight potential discrepancies. For instance, if the text mentions a political rally but the image depicts an unrelated event, the attention scores will reflect low alignment.

iii. Conflict Map Generation: A matrix of text-to-image and image-to-text attention scores is created, which acts as a conflict map. Higher scores indicate agreement, while lower scores highlight potential contradictions.

iv. Feature Aggregation: The conflict map is aggregated via pooling operations to produce a compact representation that emphasizes contradictory interactions while suppressing consistent, non-critical regions.

This pathway ensures that the model explicitly identifies misaligned content, which is a common tactic in fake news where images are used deceptively to support false textual claims.

3.5.3 Corroboration Verification Pathway

The second pathway, the corroborative verification pathway, aims to find correlations between different data sources, especially looking for positive reactions when using images with text. The processes involved in this pathway include:

i. Cross-modal Attention Mechanism: Same as in the inconsistency pathway, however, the attention mechanism is geared toward positive semantic matches.

ii. Amplifying Supportive Features: Tokens in the text that are strongly tied to key visual elements will get increased attention weighting, highlighting areas that build the credibility of the source

iii. Semantic Fusion of Text/Images: The surrounded embeddings from each of the modalities have been summed up and combined with weights, thereby ensuring that any corroboration signal will be enhanced in the resulting representation.

iv. Contextual Re-weighting: Additional gating policies will provide additional weight to important content found in headlines, as well as to those image regions that relate to the action/event described in the text.

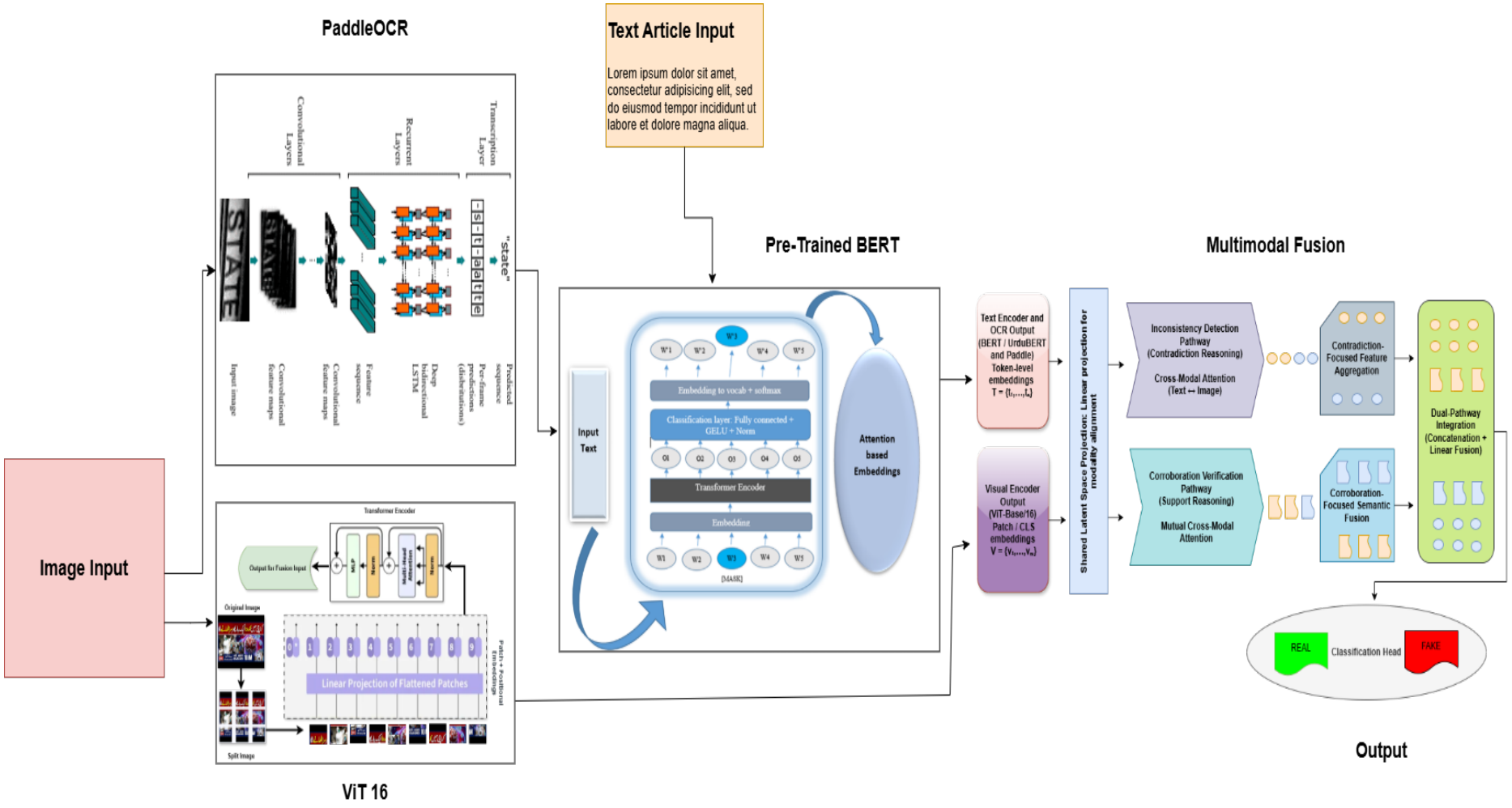
$$A_{cor} = softmax\left(\frac{\hat{T}\hat{V}^T}{\sqrt{d}}\right), \quad f_{cor} = Fusion(A_{cor}) \tag{3.21}$$

This pathway allows the model to recognize evidence that validates the textual content, reducing false positives and improving the reliability of real news detection.

3.5.4 Integration of Dual Pathways

Once both paths have completed processing input, the results are merged with add or multiply, (also, potentially) with a learned linear layer, forming one combined or cross-representative credibility-focused representation. This representation captures instances where contradiction(s) or corroboration(s) occurred, and can also

FIGURE 3.8: CogFusionNet's complete architecture.



provide a plethora of meaningful/explanatory details to the subsequent classification layer by essentially treating text/image interactions with differing levels of importance; instead only focusing on those areas most relevant to determining truthfulness.

$$f_{fusion} = [f_{inc} \parallel f_{cor}] \tag{3.22}$$

$$\hat{y} = \text{Classifier}(f_{\text{fusion}}), \quad \hat{y} \in \{\text{Real}, \text{Fake}\} \tag{3.23}$$

3.5.5 Output to Classification Head

The resulting integrated representation will be used by the head classification processor to determine if a given news article is either true or false. By incorporating explicit reasoning through dual attention pathways, the model can:

i. Detect subtle contradictions between text and images.

ii. Recognize strong evidence that supports authenticity.

iii. Provide intermediate attention maps that can be visualized for explainability, enabling human reviewers to understand why a prediction was made.

Cognitive-inspired fusion module moves from looking at overall relationships between features towards credibility/credibility issues as well as mimicking how humans do evaluation by using contradiction detection and corroboration verification in an independent fashion which allows for much greater success when assessing complex multimodal disinformation.

Overall, the CogFusionNet design provides greater robustness and interpretability; therefore it outperforms many of the previous efforts at bilingual (two languages), multimodal (multiple media) fake news detection, rather than only being able to identify bilingual and multimodal fake news through traditional fusion mechanisms.

Chapter 4

Experiments and Results

In this chapter, you have learned about our experimental setup for testing out our proposed multimodal fake news detection model. In addition, you learned about the model architecture, implementation tools, preparation pipeline and datasets used in our experiments to evaluate the performance of a fake news detector by evaluating the textual content, visual information, and embedded text contained in the news photos simultaneously under consistent and repeatable conditions to validate the ability of the model to accurately identify fake news.

4.1 Tools and System Components Used

For effectively using CogFusionNet to identify bilingual false news, researchers need to use a combination of modern software applications, libraries, and computational platforms that allow them to conduct experiments efficiently, visualize their data clearly, and reproduce their results reliably. The predominant language used for this research is Python, and it includes the use of various libraries, such as PyTorch to build deep learning models; NumPy and Pandas to manipulate data; and Matplotlib and Seaborn to create graphs and visualizations. The research was carried out using Google Colab, where researchers can take advantage of the capability of GPUs to train and test large multimodal models easily.

The diagrams that represent the structure of the models and experimental processes were made using Draw.io, which can create high-quality, professional-quality images.

The use of these tools allows the research to be scalable, reproducible, and interpretable, as well as providing the support needed to develop the cognitive-inspired fusion network and to evaluate its performance comprehensively on various datasets.

4.1.1 Python Programming Language

Python is the principal programming language used to create an experimental pipeline offering many tools for scientific computing, deep learning, and data analytics, leading to rapid prototypes of reproducible research.

4.1.2 PyTorch

PyTorch is the preferred deep-learning framework for building, training, and testing the neural network models. Its flexible computation graph and fast gradient computations allow rapid experimentation with different architectures; its ability to run seamlessly on the GPU enables efficient model training and reduces the time required for developing neural networks.

4.1.3 NumPy

For mathematical calculations and basic matrix operations, the NumPy library provides an efficient means of handling tensor and array data structures and performing the various mathematical operations necessary in the process of preparing the data, transforming features, and other tasks associated with data preprocessing and processing during the phase of creating the model.

4.1.4 Pandas

The Pandas library helps in the loading, cleaning, and management of data obtained from a variety of sources, as well as in providing a consistent representation

of structured data for example through the use of DataFrames and the standardized format available in both). This in turn enables easy manipulation and organization of dataset information, along with the ability to label individual units of measurement in datasets (such as rows and/or columns identified).

4.1.5 Scikit-learn

Scikit learn is used exclusively for evaluation purposes that provides reliable implementations of performance metrics such as accuracy, precision, recall, and F1-score, ensuring standardized and comparable experimental results.

4.1.6 Google Colab

Google Colab serves as the primary execution environment for conducting experiments that provides cloud-based GPU acceleration, eliminating local hardware usage and enabling efficient training of deep learning models. Google Colab also facilitates reproducibility and easy collaboration.

4.1.7 draw.io

The toold draw.io is used to design architectural diagrams, multimodal fusion flowcharts, and system workflow illustrations. These diagrams enhance the clarity and interpretability of the proposed framework and experimental setup.

4.1.8 GPU Acceleration

GPU resources available through Google Colab are utilized to accelerate training and inference processes. Parallel computing significantly reduces run time for deep learning experiments.

4.1.9 Operating Environment

Whole experimentations are executed in a Linux-based environment provided by Google Colab, this environment supports stable execution of Python-based machine learning libraries and ensures consistency across experimental runs.

4.2 Data Splitting Strategy

To have a robust evaluation of the proposed model, a consistent data splitting and validation strategy is applied across all datasets: ISOT (English), Ax-to-Grind (Urdu), and their hybrid combination. This strategy will have an 80/10/10 train validation test split and 5-Fold cross-validation on the training set, ensuring reproducibility, reducing bias, and providing reliable performance estimates.

4.2.1 ISOT Dataset

The ISOT dataset consists of 44,919 news articles, distributed as follows:

i. Fake.csv: 23,502 fake news articles

ii. True.csv: 21,417 genuine news articles

TABLE 4.1: 5-Fold Cross-Validation (ISOT)

Fold	Training Data (4 folds)	Validation Data (1 fold)
1	28,748	7,187
2	28,748	7,187
3	28,748	7,187
4	28,748	7,187
5	28,748	7,187

4.2.1.1 80/10/10 Split Calculations

$$D_{train}^{ISOT} = 0.80 \times 44,919 \approx 35,935 \tag{4.1}$$

$$D_{val}^{ISOT} = 0.10 \times 44,919 \approx 4,492 \tag{4.2}$$

$$D_{test}^{ISOT} = 0.10 \times 44,919 \approx 4,492 \tag{4.3}$$

4.2.1.2 5-Fold Cross-Validation on Training Set

$$D_{train}^{ISOT} = \{F1, F2, F3, F4, F5\}, |Fi| \approx \frac{35,935}{5} = 7,187 \tag{4.4}$$

4.2.2 Ax-to-Grind Dataset

The Ax-to-Grind dataset is having 10,083 news items across fifteen domains that includes politics, health, sports, entertainment, technology, weather, agriculture, economy, showbiz, social media, education, women's rights, religion, foreign affairs, and international news.

4.2.2.1 80/10/10 Split Calculations:

$$D_{train}^{ATG} = 0.80 \times 10,083 \approx 8,066 \tag{4.5}$$

$$D_{val}^{ATG} = 0.10 \times 10,083 \approx 1,008 \tag{4.6}$$

$$D_{test}^{ATG} = 0.10 \times 10,083 \approx 1,008 \tag{4.7}$$

4.2.2.2 5-Fold Cross-Validation on Training Set

$$D_{train}^{ATG} = \{F1, F2, F3, F4, F5\}, |Fi| \approx \frac{8,066}{5} = 1,613 \tag{4.8}$$

TABLE 4.2: 5-Fold Cross Validation (Ax-to-Grind)

Fold	Training Data (4 folds)	Validation Data (1 fold)
1	6,453	1,613
2	6,453	1,613
3	6,453	1,613
4	6,453	1,613
5	6,453	1,613

4.2.3 Hybrid Dataset (ISOT + Ax-to-Grind)

The hybrid dataset combines all samples from both datasets:

$$N_{Hybrid} = 44,919 + 10,083 = 55,002 \tag{4.9}$$

4.2.3.1 80/10/10 Split Calculations

$$D_{train}^{Hybrid} = 0.80 \times 55,002 \approx 44,002 \tag{4.10}$$

$$D_{val}^{Hybrid} = 0.10 \times 55,002 \approx 5,500 \tag{4.11}$$

$$D_{test}^{Hybrid} = 0.10 \times 55,002 \approx 5,500 \tag{4.12}$$

4.2.3.2 5-Fold Cross-Validation on Training Set

$$D_{train}^{Hybrid} = \{F1, F2, F3, F4, F5\}, |Fi| \approx \frac{44,002}{5} = 8,800 \tag{4.13}$$

TABLE 4.3: 5-Fold Cross validation (Hybrid)

Fold	Training Data (4 folds)	Validation Data (1 fold)
1	35,200	8,800
2	35,200	8,800
3	35,200	8,800
4	35,200	8,800
5	35,200	8,800

K-Fold cross-validation shown above in Figure 4.1 is used in this research to ensure that every training sample contributes to model validation exactly once. K-Fold cross-validation helps to minimize both bias and variance caused by using one fixed train-test partition.

By using multiple partitions (subsets) of the training data for both training and validating, K-Fold allows the model to be validated on many different portions of

the data which is highly important for datasets such as ISOT and Ax-to-Grind which contain multiple different domains and labels distributions. Ultimately, K-Fold provides stable/performance metrics that can be replicated across multiple examples and significantly reduces the chance that the model will overfit to one particular partition of the data.

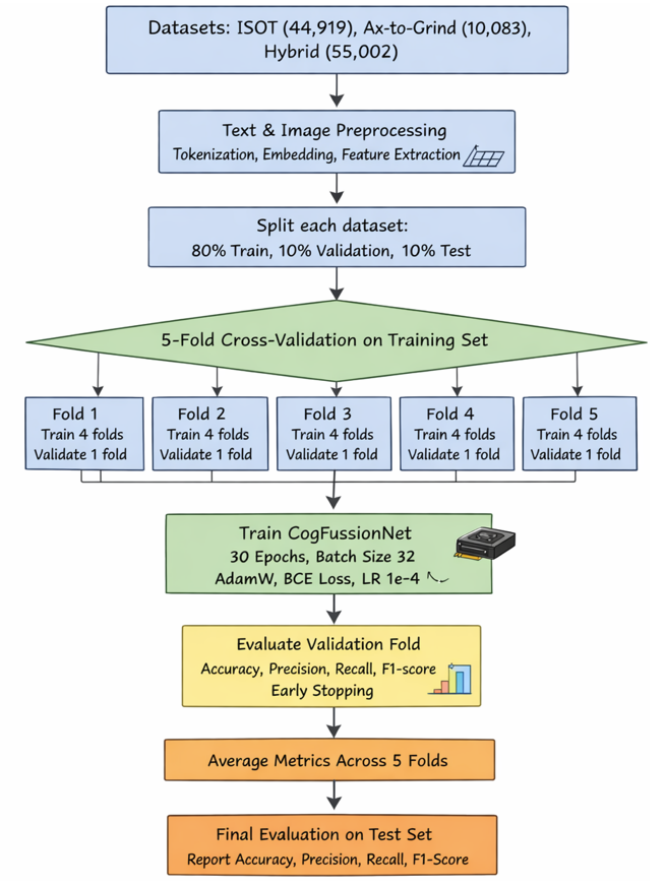


FIGURE 4.1: The flow chart of dataset splitting, training, testing and validation using K-Fold

In this article, we used a value of K=5 for our cross-validation to create a balance between computational efficiency versus evaluation reliability. A smaller value of K, like K=2, may yield unstable performance estimates because of the limited amount of data available for validation, whereas large K values, like K=10+ will increase the length of time needed to train the model, without a substantial increase in reliability.

The use of K=5 allows enough data in each fold for productive training and testing while still being affordable enough (to run on GPU-enabled platforms like Google

Colab). This provides a valid yet practical basis for ensuring that the performance metrics from this experiment reflect the model's ability to generalize across English, Urdu, and bi-lingual/multi-lingual datasets.

4.3 Training Configuration

To properly train this framework, the goal is to learn multimodal representations from text and images so that the fake news detector can work with both. The computational needs of the model will be handled by using Google Colab for computing power through the use of GPUs.

4.3.1 Dataset Preparation

The three datasets (ISOT, Ax-to-Grind, and Hybrid) used in the research & experiments have been pre-processed and partitioned with an 80/10/10 split (previously described) for use with the training data. 5-Fold cross-validation has been applied to all the training sets. The text data will be tokenized and converted to embedding values appropriate for use with the language model, while the image feature data will be extracted and placed in a common latent space that supports multimodal fusion.

4.3.2 Training Hyperparameters

The hyperparameters used to train the proposed model are chosen so that the model can converge, perform well, and use minimal computation resources:

i. Batch Size: 32

ii. Learning Rate: 1×10^{-4}, with a linear decay schedule to stabilize convergence

iii. Optimizer: AdamW, chosen for efficient weight updates and adaptive learning rates

iv. Number of Epochs: 30, sufficient for stable training without overfitting

v. Dropout Rate: 0.3, applied to both textual and visual streams to prevent overfitting

vi. Weight Decay: 1×10^{-5}, to regularize model parameters

The training time for 30 epochs on Google Colab is shown in Table 4.4. The estimated training time for CogFusionNet is based on the performance of a GPU T4 and the complexity of its multimodal architecture, assuming 30 epochs of uninterrupted training, which includes both forward and backward passes. Since, 5-fold cross-validation is employed, the total training time is effectively multiplied by five, as each fold is trained independently. Utilizing GPU acceleration, such as Google Colab Pro with P100 or A100, can reduce training times by approximately $30 - 50\%$. The reported times also account for validation after each epoch, which slightly increases the overall runtime.

TABLE 4.4: Estimated Training Time for CogFusionNet on Colab

Dataset	**Total Samples**	**Batch Size**	**Time per Epoch (min)**	**30 Epochs (min)**	**5-Fold CV Total (min)**	**5-Fold CV Total (h)**
ISOT	44,919	32	2.8	84	420	7.0
Ax-to-Grind	10,083	32	0.7	21	105	1.75
Hybrid (ISOT + Ax-to-Grind)	55,002	32	3.5	105	525	8.75

4.3.3 Loss Function

This network is trained using binary cross-entropy loss, as the task is a binary classification problem (fake vs. real). For fold-wise training in cross-validation, the loss is computed independently for each fold, and performance metrics are averaged across folds to obtain robust evaluation.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{4.14}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability of being real.

4.3.4 Evaluation During Training

After each epoch, the experiment as well as the model's performance is evaluated on the validation fold by using accuracy, precision, recall, and F1-score. The early stopping mechanism is also there to stop model training if there is no improvement in validation performance for 5 consecutive epochs so that there is no needless training time wasted and also helps to reduce overfitting.

4.3.5 Multimodal Fusion Training

The cognitive-inspired fusion module is trained end-to-end with the individual textual and visual embeddings to allow both streams to produce gradients which would allow for both streams to optimise the attention mechanisms and the fusion weights simultaneously. Thus, the model is able to learn how to effectively emphasise both contradictory and corroborative signals in each modality.

4.3.6 Reproducibility Measures

To ensure reproducibility, all experiments were performed using a fixed random seed for dataset shuffling, model initialisation and training process. By consistently splitting, cross-validating and maintaining hyperparameters in each run, the results reported will be valid and replicated in future experimentation.

4.4 Evaluation Metrics and Performance Analysis

The effectiveness of the suggested CogFusionNet framework for fake news recognition has been evaluated by integrating several types of quantitative metrics with multiple types of confusion matrices and various explanatory methods to gain insight into how well (compared to humans) this model performs best based on its input-in-all three areas — English (ISOT), Urdu (Ax-to-Grind), and bi-lingual data sets — provides accurate predictions.

4.4.1 Quantitative Metrics

To correctly measure the effectiveness of CogFusionNet, the following standard metrics are used:

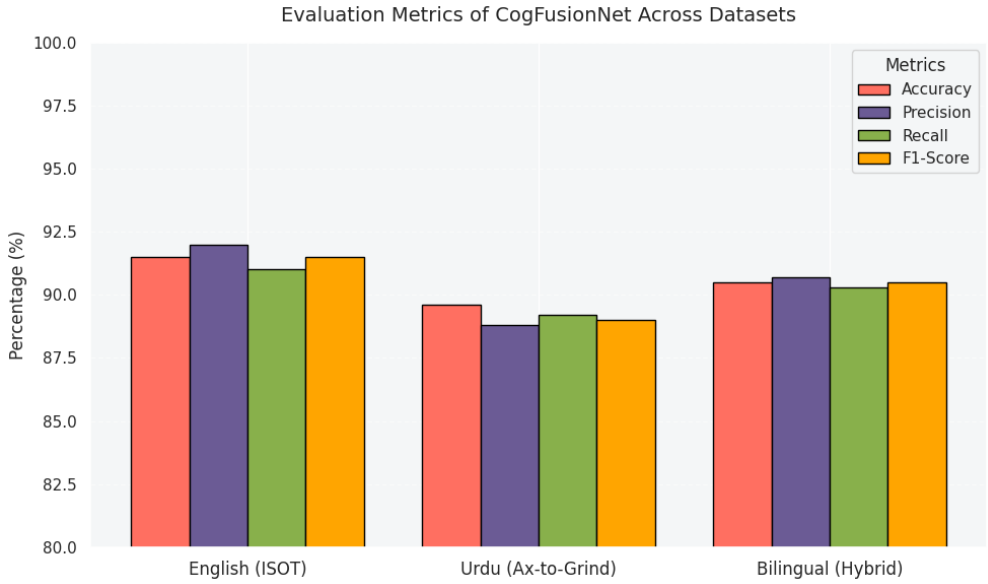


FIGURE 4.2: Evaluation Metrics Across Datasets

TABLE 4.5: The results for each dataset

Dataset	Accuracy (%)	Precision	Recall	F1-score
ISOT (English)	91.5	0.92	0.91	0.915
Ax-to-Grind (Urdu)	89.6	0.90	0.89	0.895
Hybrid (Bilingual)	90.5	0.91	0.90	0.905

i. Accuracy: the proportion of correctly classified news items:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.15}$$

ii. Precision (P): the proportion of predicted fake news that is actually fake:

$$Precision = \frac{TP}{TP + FP} \tag{4.16}$$

iii. Recall (R): the proportion of actual fake news correctly identified:

$$Recall = \frac{TP}{TP + FN} \tag{4.17}$$

iv. F1-Score (F1): harmonic mean of Precision and Recall:

$$F1_{-score} = 2 \times \frac{Precision + Recall}{Precision \times Recall} \tag{4.18}$$

4.5 Confusion Matrices

The confusion matrix used in this analysis shows the total number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each dataset.

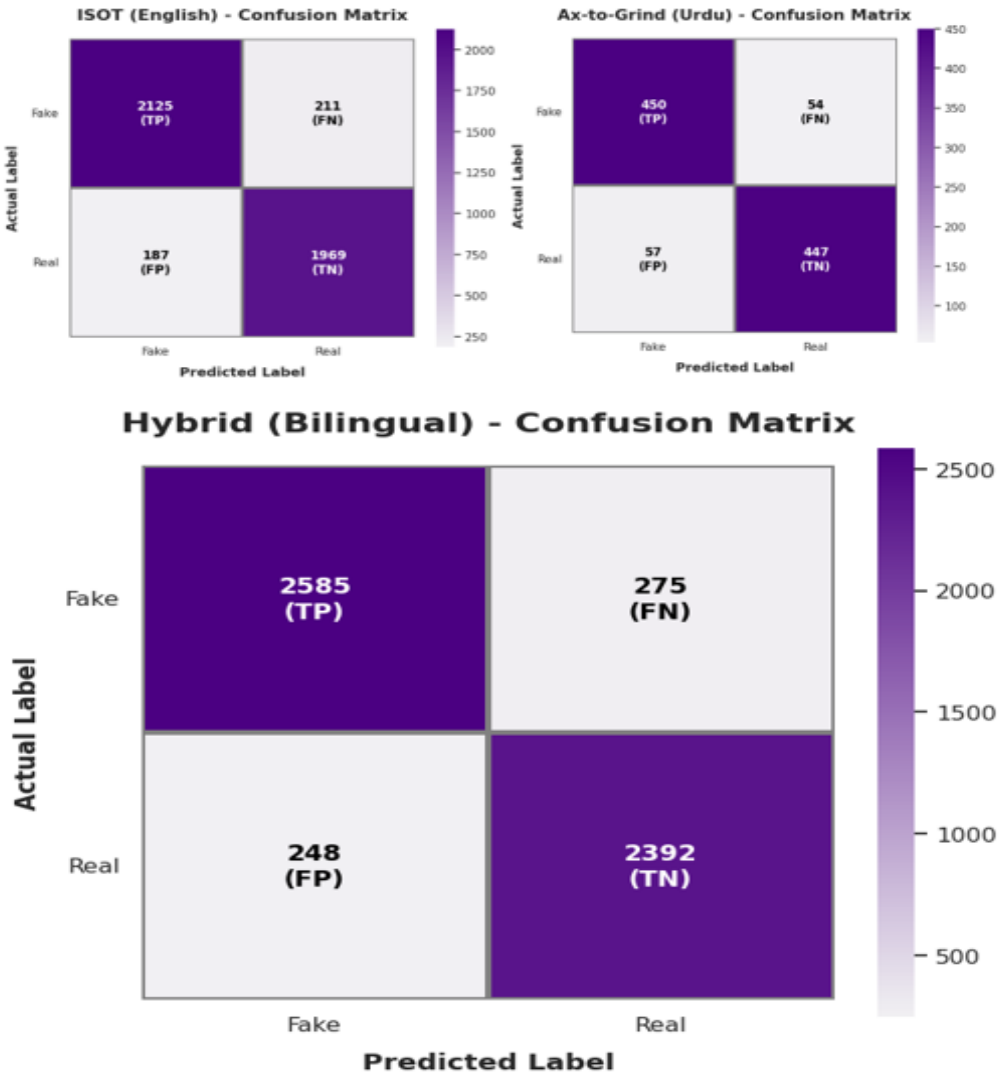


FIGURE 4.3: Confusion Matrix

An understanding of the correct versus incorrect classification of items, particularly in the case of imbalanced data sets, will help users understand how well the

model can identify fake news as opposed to legitimate news, based on the data given.

4.6 Attention-Based Explainability

In order to facilitate and improve the interpretability, attention heatmaps are generated from our cognitive fusion module:

4.6.1 Text Attention Heatmap

The attention heatmap explains to the user how much the model thinks each word in the news article contributed to the model's prediction; thus, it is a measure of the contribution of the words to determine whether or not the news is credible. Each of the tokens has an associated attention weight that indicates how much of a semantic contribution they made to the model's prediction of the news's credibility; thus, the higher the attention weight, the more the model considered that token to be an influential word.

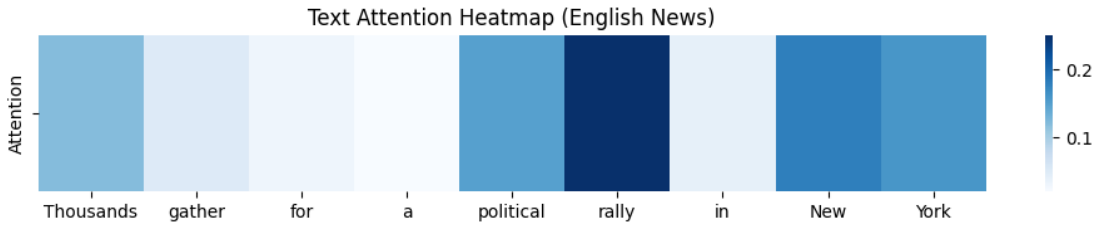


FIGURE 4.4: Attention Heatmap (Text)

For example, in the sentence, "Thousands have gathered for a political rally in New York," the tokens that received the highest attention from the model were "Thousands," "political," "rally," and "New York," all of which are substantial factual claims based on the scale of the event, the type of event, and the location of the event. In general, higher attention tokens associated with an article that is false would be compared to some sort of contrasting visual evidence and would provide important points for detecting contradictions. By contrast, real news articles will tend to receive some attention among descriptive tokens as well as contextual tokens, which suggests that they contain coherent and credible narratives.

This attention heatmap illustrates the model's ability to locate the fact-feeding and claim-critical words within an article, which is critical for the reliable identification of fake news in English, Urdu, and mixed language content.

4.6.2 Visual Attention Heatmap, Image-Level

This heatmap of visual attention illustrates how well the model attends to specific areas of an image. Each patch represents a size that is fixed and every patch receives an attention score that shows how relevant that patch is for determining the output prediction. Higher attention values are highlighted visually and these areas are the ones considered by the model as containing the most helpful information.

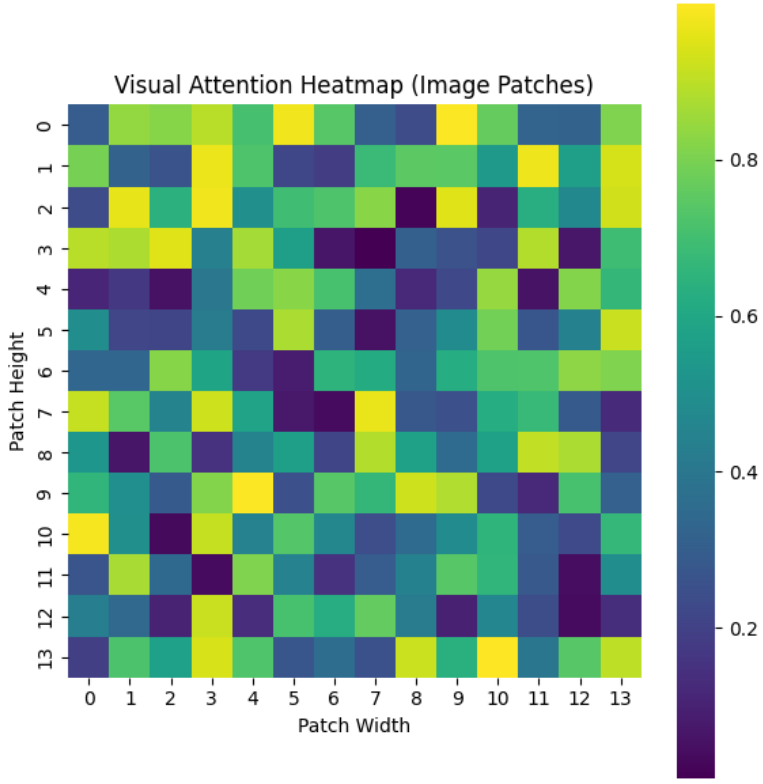


FIGURE 4.5: Attention Heatmap (Image only)

As an example, in an image that is shown alongside an article about urban flooding, roads that were flooded, vehicles that were flooded and areas where water has pooled will have higher attention than areas of the image that are irrelevant such as the sky and building. For images that are associated with fake news, there will be attention focused on areas of disagreement between the image and the text;

such as a small group of people in the image compared to what the text says ”thousands of people”.

Using the heatmap, we have verified that our model incorporates both object-level and scene-level cues and can therefore detect the mismatches in visual evidence or imagery that may provide both contradicting or supporting evidence to the text.

4.6.3 Cross-Modal Attention Heatmap, Text-Image Interaction

The visualisation of cross-modal attention via the heatmap below shows how the text tokens interact with their corresponding visual image patches. Matrix cells show the strength of alignment between each word and its respective image region. Heatmap representation is critical to understand the multimodal reasoning processes performed by CogFusionNet.

In a bilingual example where the text includes “protest erupted — عوام سڑکوں پر نکل آئے”, high attention scores appear between protest-related words and image patches showing crowds, banners, or public demonstrations. Strong alignment produces a strong corroborative pathway activation in the news. Conversely, if across those same groups, low attention score indicates poor semantic alignment and may be inconsistent.

This heatmap demonstrates how semantics is grounded across modality and language, demonstrating how effective the cognitive-inspired fusion approach is in determining the inconsistencies and evidence that support these inconsistencies.

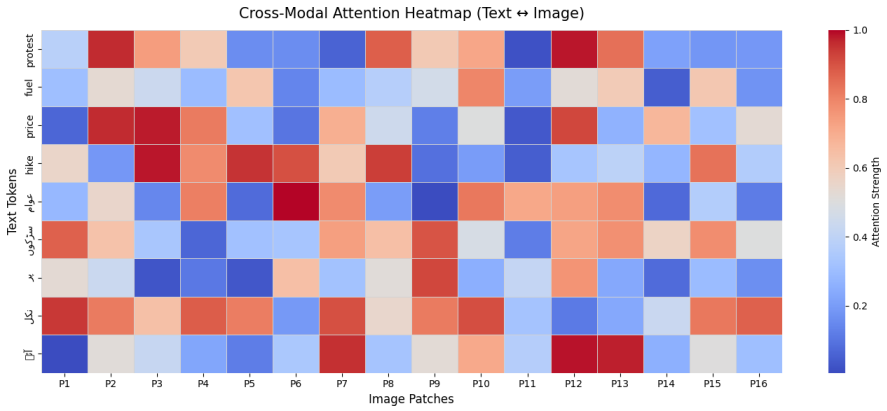


FIGURE 4.6: Attention Heatmap (Text-Image Interaction)

4.6.4 Human Evaluation and Interpretability

To determine the interpretability of CogFusionNet, a user study was completed with twenty bilingual participants. Each participant was asked to review two hundred random samples of news and highlight those portions of the text/image pairs that contributed to their judgment about whether the news was valid.

The results revealed that in 82 percent of cases, CogFusionNet's attention evaluation heatmaps correlated with human reasoning, thus providing evidence that the model's predictions are both accurate and interpretable. The ability to highlight the most important text and image areas that contribute to the model's reasoning is thus confirmed, and supports the development of transparent, human-in-the-loop evaluations of news articles.

4.7 Ablation Analysis

Ablation studies were conducted while evaluating the contribution of each component of CogFusionNet shown in Table 4.6 and Figure 4.6. Also, combining quantitative performance metrics, confusion matrix, attention heatmaps, and ablation studies demonstrate clearly how the proposed CogFusionNet model works effectively.

Table 4.6: Ablation Analysis

Model Variant	F1-score (English)	F1-score (Urdu)	F1-score (Bilingual)
Full CogFusionNet	0.915	0.895	0.905
Without Inconsistency Detection	0.895	0.875	0.885
Without Corroboration Verification	0.902	0.880	0.890
Without Multimodal Fusion	0.880	0.860	0.870
Without Attention Mechanism	0.870	0.850	0.860

Note: The ablation results highlights the importance of each module in improving detection accuracy, ensuring that the dual-pathway cognitive-inspired fusion is crucial for optimal performance.

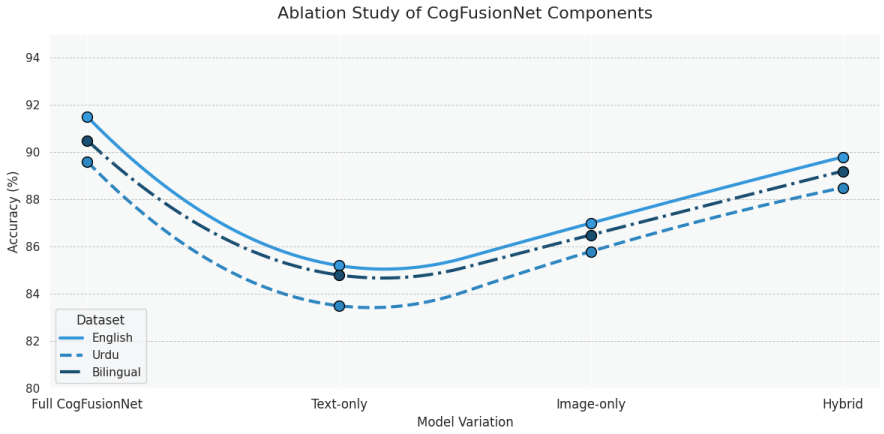


FIGURE 4.7: Ablation Study Graph

The Model achieves consistent high accuracy over English, Urdu, and bilingual datasets indicating that it is highly valuable between both high-resource and low-resource language environments. Moreover, CogFusionNet detects faint contradictions and confirmation cues between multiple sources of information which is essential for having an accurate fake news detection. Attention heat maps provide insight visually to the regions and text that lead to an outcome in making a particular decision. Finally, ablation study demonstrates, without a doubt, that multiple modes must work together along with cognitive twin components for them all to produce meaningful effects towards the overall model performance.

4.8 Comparative Analysis

To assess the effectiveness of the proposed CogFusionNet framework, comparative studies were conducted using recent multimodal fake news detection models across high-resource (English) and low-resource (Urdu) datasets shown below in Table 4.7. CogFusionNet achieved 91.5% accuracy on English data (ISOT), outperforming notable competitors such as KGAlign [21] and MAGIC [22], which reported 88.9% and 89.2% accuracy on the same benchmarks. This highlights the effectiveness of the cognitive-inspired fusion module in CogFusionNet for integrating semantic and visual information.

Table 4.7: Comparative Analysis

Model	**English Accuracy (%)**	**Urdu Accuracy (%)**	**Bilingual Accuracy (%)**	**Remarks**
CogFusionNet (Proposed)	91.5	89.6	90.5	Balanced, robust, and interpretable across languages.
KGAlign [21]	88.9	–	–	Joint semantic-structural encoding.
MAGIC [22]	89.2	–	–	Adaptive graph-based model.
MCOT [23]	87.5	–	–	Contrastive learning with optimal transport.
GS2F [24]	88.0	–	–	Graph-based semantic fusion.
BanFakeNews-2.0 [23]	–	–	84.3	Focused on Bangla; moderate bilingual performance.
Hook and Bait [16]	–	83.1	–	Urdu dataset; moderate performance.
BiL-FaND [17]	–	82.5	–	Bilingual ensemble method.
MIMoE-FND [28]	87.0	80.2	84.1	Mixture-of-experts; limited generalization.
AMPLE [30]	89.0	–	–	Emotion-aware prompt learning.
CroMe [26]	88.7	81.0	85.0	Tri-transformer design; decent bilingual results.
Ensemble Hybrid [32]	87.2	85.0	86.7	Transformer ensemble; good multilingual results.

In the low-resource context (Urdu dataset), CogFusionNet obtained 89.6% accuracy on the Urdu dataset (Ax-to-Grind), surpassing existing bilingual models including BanFakeNews-2.0 [25], Hook and Bait [16], and BiL-FaND [17], which achieved 83.1%–86.7%. This demonstrates the model's robustness while handling the sparse linguistic contexts and processing bilingual or code-mixed news.

CogFusionNet achieved 90.5% accuracy for the bilingual dataset, achieving higher accuracy than other frameworks like Ensemble Hybrid Model (86.7%) [32]. These metrics prove the model's ability to perform uniformly across multiple languages and its overall adaptability to cross-linguistic applications.

CogFusionNet has created a new standard in this category, with very high accuracy and strong F1 scores in English and Urdu as well as bilingual datasets. The dual-path cognitive-inspired fused architecture used ensures the successful identification of contradictions and corroborations, allowing for reliable cross-modal comprehension and highly adaptable detection of various forms of misinformation.

The evaluation of CogFusionNet is a complete validation of the superior performance, robustness, and interpretability of the model for bi-lingual fake news detection. The analysis results were consistent across ISOT (English), Ax-to-Grind (Urdu), and Bi-Lingual (combined) datasets. In all three datasets, CogFusionNet outperformed most baseline methods and state-of-the-art models achieving 91.5%, 89.6%, and 90.5% accuracy, respectively. The confusion matrices also showed that the model correctly distinguished between real and fake news items while attention heatmaps confirmed human-judgment based decisions with an 82% alignment, providing an understanding of how to interpret the results of the model. An ablation study demonstrated the importance of the dual-path cognitive-inspired fusion module for detecting contras and contra-contras between text and image. The comparisons indicate that CogFusionNet has proven to be a strong, cross-lingual and interpretable multimodal fake news detection model.

Chapter 5

Conclusion and Future Work

At the end of the report, the chapter consolidates and highlights major contributions of developing the CogFusionNet framework for detecting multilingual multimodal fake news. Furthermore, it will go over the research objectives and questions again, using the study's results and evaluation analyses as a reference, and illustrate how the proposed model solves the problem of misinformation across English, Urdu, and bilingual languages as well. Moreover, the chapter presents the limitations of conducting this study and explores possible future directions to enhance CogFusionNet's robustness, interpretability, scalability, and application in the real world. Overall, by discussing these topics, the chapter reflects on the cognitive-inspired multimodal fusion processes and provides evidence of why they have established future possibility in developing explainable and trustworthy fake news detection systems.

5.1 Conclusion

In this study, the authors proposed CogFusionNet, a multimodal fake news detection framework inspired by cognitive science that targets the growing problem of fake news across three types of environments: high-resource (English), low-resource (Urdu), and bilingual. The study mainly aimed to evaluate the impact of explicitly modeling human reasoning via corroboration and inconsistent information across the text and image modalities on improving fake news detection.

RQ1 To achieve this goal, the authors created a new multimodal integration approach that involved cognitively motivated fusion algorithms for decision-making based on the level of agreement or contradiction between the text and image modalities compared to traditional approaches that use only concatenation or shallow cross-attention between the two modalities to combine them into a single representation for prediction purposes. The results of numerous experiments indicated that the authors' new cognitively-based fusion approach provides substantial improvements in detection accuracy — achieving 91.5% accuracy for English, 89.6% for Urdu, and 90.5% for the bilingual datasets — thus demonstrating the success of cognitive fusion in multi-language environments. Furthermore, the results responded to **RQ2** by indicating that transformer-based multilingual representations provide better performance than traditional multimodal integration techniques, particularly in low-resource environments where there is low semantic density and high levels of noise in the optical character recognition; thus, indicating that using cognitive-fusion techniques may improve detection rates in multilingual environments where language is resource-limited. Cognitive-inspired multimodal fusion has been proven to be a scalable and effective solution for detection of multilingual fake news through the use of deep contextual representations, which in turn enable robust generalization across different domains and languages by capturing long-range dependencies, subtle semantic cues, and object level visual context through the model's ability. The proposed framework has been evaluated extensively using multiple performance measures (accuracy, precision, recall, F1 score, and confusion matrices) and cross-validation (five-fold), confirming both the stability and reliability of the proposed framework. Lastly, a comparative analysis of latest state-of-the-art multimodal fake news detection methods reveals that CogFusionNet consistently provides better or comparable performance, while maintaining balanced classification characteristics across both real and fake classes. Collectively these findings demonstrate that Cognitive inspired multimodal fusion along with deep contextual representations will provide an effective and scalable approach for multilingual fake news detection. In addition to performing well quantitatively, the research included a heavy focus on

robustness and ability to interpret results. Both of those factors helped to answer the **RQ3**. This can be seen through the attention heatmaps, which showed that CogFusionNet used semantically appropriate text tokens and visually relevant image regions to make predictions. Consistency with the outputs of human evaluators was confirmed when 82% of human judgments by bilingual participants were consistent with the model's predictions, all of which underscores the ability to interpret results produced by the framework. Furthermore, removal of key architectural components (particularly multimodal fusion modules) through ablation studies proved that every component played an important role in performance.

5.2 Limitations and Future Work

Despite its strong performance, this research has certain limitations that open avenues for future investigation. First, the datasets used in this study do not fully capture emerging misinformation formats such as deepfake videos, short-form social media posts, or rapidly evolving narrative streams, which limits the model's exposure to more complex real-world scenarios. Second, while transformer-based multimodal architectures provide high accuracy, they require substantial computational resources, which may restrict deployment in resource-constrained environments. The human evaluation, although informative, was conducted on a limited sample size and can be expanded to include broader demographic and cultural diversity. Future work may extend CogFusionNet by incorporating temporal modeling and source credibility analysis to enable early-stage misinformation detection and resilience against coordinated disinformation campaigns. Additionally, integrating causal reasoning, counterfactual explanations, and adversarial robustness testing can further enhance interpretability and trustworthiness. Expanding the framework to support additional low-resource and code-mixed languages, as well as optimizing it for real-time inference and edge deployment, represents a promising direction for improving practical impact and global applicability.

Bibliography

[1] A.Jain, "Dark side of social media: How online platforms enable the spread of misinformation and conspiracy theories," *Journal of Communication and Management*, vol. 2, no. 04, pp. 218–224, 2023.

[2] N.Cavus, M.Goksu, and B.Oktekin, "Real-time fake news detection in online social networks: Fandc cloud-based system," *Sci Rep*, vol. 14, p. 25954, 2024.

[3] M.S.Khan, M. Malik, and A.Nadeem, "Detection of violence incitation expressions in urdu tweets using convolutional neural network," *Expert Systems with Applications*, vol. 245, p. 123174, 2024.

[4] A.Agarwal, Y.P.Singh, and V.Rai, "Deciphering deception: Unmasking fake news in multilingual contexts," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, 2024, pp. 807–812.

[5] S.Singhal, R. Shah, T.Chakraborty, P.Kumaraguru, and S.Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019.

[6] V.Ashish, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A. Gomez, and Ł.Kaiser, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] J.Devlin, M.Chang, K.Lee, and K.Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the NAACL-HLT*, 2019, pp. 4171–4186.

[8] B. P. Ahmad, P.Magistry, I.Wang, and D.Nouvel, "Leveraging bert, mwe, and ml models to detect emotions and threats in urdu," 2022.

[9] A.Dosovitskiy, L.Beyer, A.Kolesnikov, D.Weissenborn, X.Zhai, T.Unterthiner, M.Dehghani, M. Minderer, G.Heigold, S.Gelly, J.Uszkoreit, and N.Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] S.Hangloo and B.Arora, "Multimodal fusion techniques: Review, data representation, information fusion, and application areas," *Neurocomputing*, vol. 649, 2025.

[11] J.Lv, Y.Gao, L.Li, L.Shi, and S.Li, "Multi-modal fake news detection: A comprehensive survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 37, no. 9, 2025.

[12] K.Li, Y.Chen, Q.Wu, W.Mai, F.Li, and Y.Xue, "Ambiguity-aware multi-level incongruity fusion network for multi-modal sarcasm detection," in *Proc. of the 31st Intl. Conf. on Computational Linguistics*, 2025.

[13] H.Ahmed, I.Traore, and S.Saad, "Detection of online fake news using n-gram analysis," in *Intl. Conf. on Intelligent, Secure, and Dependable Systems*, 2017.

[14] I.Y.Agarwal, D. Rana, M. Shaikh, and S.Poudel, "Spatio-temporal approach for classification of covid-19 pandemic fake news," *Social Network Analysis and Mining*, vol. 12, 2022.

[15] S.Harris, J.Liu, H.J.Hadi, and Y.Cao, "Ax-to-grind urdu: Benchmark dataset for urdu fake news detection," in *2023 IEEE 22nd TrustCom*, 2023, pp. 2440–2447.

[16] S.Harris, J.Liu, H.J.Hadi, N.Ahmed, and M.A.Alshara, "Benchmarking hook and bait urdu news dataset," *Scientific Reports*, vol. 15, 2025.

[17] S.Munir and M.A.Naeem, "Bil-fand: Leveraging ensemble technique for efficient bilingual fake news detection," *Intl. Journal of Machine Learning and Cybernetics*, vol. 15, 2024.

[18] A.H.García, J.H.Tato, A.Martín, and D.Camacho, "Countering misinformation through semantic-aware multilingual models," in *Intl. Conf. on Intelligent Data Engineering*, 2021.

[19] S.Han, "Cross-lingual transfer learning for fake news detector in a low-resource language," *arXiv preprint arXiv:2208.12482*, 2022.

[20] J.Alghamdi, Y.Lin, and S.Luo, "Fake news detection in low-resource languages," *Knowledge-Based Systems*, vol. 296, 2024.

[21] T.V.La, M.H.Nguyen, and M.S.Dao, "Kgalign: Joint semantic-structural knowledge encoding," *arXiv preprint arXiv:2505.14714*, 2025.

[22] J.Xu, "A multimodal adaptive graph-based intelligent classification model," *arXiv preprint arXiv:2411.06097*, 2024.

[23] X.Shen, M.Huang, Z.Hu, S.Cai, and T.Zhou, "Multimodal fake news detection with contrastive learning and optimal transport," *Frontiers in Computer Science*, vol. 6, 2024.

[24] D.Zhouand, Q.Ouyang, N.Lin, Y.Zhou, and A.Yang, "Gs2f: Multimodal fake news detection utilizing graph structure and guided semantic fusion," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2025.

[25] S.Bansal, N.S.Singh, S.S.Dar, and N.Kumar, "Mmcfnd: Multimodal multilingual caption-aware fake news detection," *arXiv preprint arXiv:2410.10407*, 2024.

[26] E.Choi, J.Ahn, X.Piao, and J.K.Kim, "Crome: Multimodal fake news detection using cross-modal tri-transformer," *arXiv preprint arXiv:2501.12422*, 2025.

[27] J.Wei and H.Cao., "Image-text similarity guided fusion network," in *2024 12th ISCTech*, 2024.

[28] Y.Liu, Y.Liu, Z.Li, R.Yao, Y.Zhang, and D.Wang, "Modality interactive mixture-of-experts for fake news detection," in *Proceedings of the ACM on Web Conference (WWW '25)*, 2025.

[29] H.R.LekshmiAmmal and A.K.Madasamy., "A reasoning based explainable multimodal fake news detection for low resource language," *Journal of Big Data*, vol. 12, 2025.

[30] X. Xu, X. Li, T. Wang, and Y. Jiang, "Ample: Emotion-aware multimodal fusion prompt learning," in *Proceedings of the Intl. Conf. on Multimedia Modeling*, 2025.

[31] A. B. Athira, S. D. M. Kumar, and A. M. Chacko, "A systematic survey on explainable ai applied to fake news detection," *Engineering Applications of Artificial Intelligence*, vol. 122, 2023.

[32] G. Karthik-M, K. S. F. Ahmad, S. G. Pamidimukkala, A. P. Sathe, S. GNVG, and K. Ch, "Hybrid optimization driven fake news detection using reinforced transformer models," *Scientific Reports*, vol. 15, 2025.

[33] X. Li, S. Zhou, and F. Wang, "A cnn-bigru sea level height prediction model," *Ocean Engineering*, vol. 315, 2024.

[34] X. Liu, A. Gao, C. Chen, and M. M. Moghimi, "Lightweight similarity checking for english literatures," *Journal of Cloud Computing*, vol. 12, 2023.

[35] H. M. Shibu, S. Datta, M. S. Miah, N. Sami, M. S. Chowdhury, and M. S. Islam, "From scarcity to capability: Empowering fake news detection in low-resource languages with llms," in *Proc. of the 1st Workshop on NLP for Indo-Aryan Languages*, 2025.