FORENSIC METHODS TO DETECT MANIPULATED NEWS MEDIA

by

Abhishek Kumar

August 2023

A thesis submitted to the

Faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfillment of the requirements for the

degree of

Master of Science

Department of Computer Science & Engineering

Abstract

In the current era, the internet has made it incredibly easy to access information from various sources. From reputable news agencies to independent reporters and even unknown individuals with extreme views, a plethora of information is available online. Social media platforms have become increasingly important in our daily lives, allowing users to stay up-to-date with the latest news and developments, share links to interesting or important information, and express their opinions.

However, abundant online information can also create problems for those trying to consume it. It can be difficult for users to distinguish between real and fake news and to identify manipulated information. This is especially true since many users may not be aware that the information they come across is false or manipulated. They may not have the time or resources to fact-check everything they read online. Given the sheer amount of information being created and shared daily, it is simply not feasible for journalists or other professionals to manually fact-check every information published online. This is where automatic forensic methods that detect inconsistencies and manipulations in news articles can be handy. By quickly identifying and exposing misleading information to the public, such methods can help prevent the spread of manipulated news and ensure that people can access accurate and reliable information.

Currently, the detection of manipulated news media is predominantly carried out through machine learning models that rely on computational power to identify manipulations automatically. Some researchers focus on extracting the semantic and contextual meaning from news articles, statements, and social media posts, which try to identify manipulated information in the news by analyzing the articles' differences in writing style and semantic meaning. On the other hand, other researchers have explored using information from social networks to detect manipulations more accurately. These methods aim to distinguish between tampered and pristine news by examining the spreading patterns of news and the statistical information related to users who engage with the propagated news.

In this thesis, we propose forensic methods for manipulated news media detection involving textual and visual features that can be extracted from news articles. Specifically, our algorithms can process Multi-Media Assets (MMAs) containing images and texts and identify image-text inconsistencies or object manipulations in images using extracted features and embedding. We aim to detect and characterize potentially tampered information in news media that further helps analysts to determine the intent and tactic for creating the manipulations and its broader objectives.

In the first chapter, we propose methods that identify parts of images manipulated by specialized techniques and localize and label them within the image. We first conduct experiments with several low-level image artifacts to identify the features that can help detect manipulations. We extensively experimented with JPEG compression levels, noise distributions, and edge/boundary artifacts in tampered and pristine images. We also combine features to get better representations for our models. We use publicly available tampering detection datasets, e.g., CASIA v1/v2, etc., for training and experimentation and present our results.

In chapter two, we extend our method with a model that detects inconsistencies within infographic images and their associated text. Infographic images contain unstructured text data with words, numbers, and a plot or graphic representation. The numbers or text can easily be manipulated to create misleading data. Thus we effectively try to identify the mismatch between data in the infographic and its associated text. Due to the lack of openly available news datasets for infographics, we create a novel dataset (MMA-infographic) with infographic image-text pairs to develop and test our methods. We leverage several natural language processing models and libraries to design our dataset. We conduct experiments on this dataset for our inconsistency detection task and present our results and findings. The experiment results show that our novel methods can identify inconsistencies in news media.

Overall, this thesis demonstrates methods that detect manipulations and inconsistencies in multi-modal media using visual and linguistic features can effectively detect manipulated news and combat the spread of false information in online media. It is hoped that a more comprehensive and reliable mthod can be developed by combining techniques to combat the spread of manipulated news.

Acknowledgments

I want to express my heartfelt gratitude to the many individuals who have contributed to the successful completion of this thesis. First and foremost, I am deeply indebted to my thesis advisor, Dr. David Doermann, for his invaluable guidance, support, and encouragement throughout the research process. His expertise, insights, and feedback were instrumental in shaping the direction and quality of this study. I could not have asked for a better mentor, and I am genuinely grateful for his unwavering commitment to my academic and personal growth.

I am grateful to University at Buffalo for allowing me to pursue graduate studies and offering the resources and facilities necessary to conduct this research. The stimulating academic environment and diverse community of researchers and students at the University at Buffalo have broadened my horizons and enriched my learning experience. I am grateful for the many opportunities I have had to collaborate with and learn from fellow graduate students, faculty, and staff members.

Finally, I wish to acknowledge the unwavering support and encouragement of my family, friends, and loved ones. Their love, understanding, and patience motivated me and helped me overcome the challenges and obstacles I encountered. Their words of encouragement, gestures of kindness, and expressions of pride have sustained me through the ups and downs of graduate school. I am lucky to have

such wonderful people in my life, and I am forever grateful for their support and love.

Thank you all for your invaluable contributions to this thesis. I could not have accomplished this milestone without your help, guidance, and encouragement. I hope this thesis will be a testament to the power of collaboration, curiosity, and perseverance, and it will inspire others to pursue their intellectual and professional goals with passion and dedication.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Background

Information is a broad term that refers to knowledge or data that has been communicated or received. It can be generated by organizing, interpreting, and transmitting data and may take many forms, including facts, figures, statistics, ideas, and opinions.

In the digital age, information has become more widely available and accessible than ever before. The ease of producing information has contributed to the increasing popularity of online details created in the past decade and made the internet increasingly important for information consumers. The internet and other technologies have made it possible to store and transmit vast amounts of information across great distances and at lightning-fast speeds. This has led to an explosion of information, with new data and knowledge being generated and shared daily.

News is a type of information specifically focused on events and developments and serves many functions, one of which is to provide relevant and helpful in-

formation to the public. News plays a critical role in shaping public opinion and attitudes. News coverage can influence how people perceive specific issues and events; for example, news coverage of a natural disaster can help mobilize public support and resources for the affected areas, while news coverage of a political scandal can damage a politician's reputation and fitness for office.

Information in the news is essential to everyone because it is a vital source of information, accountability, and influence. But it is important to remember that not all information is created equal. Some data may be accurate and reliable, while others may be misleading, biased, or false. The massive amount of information published online is challenging and time-consuming to verify. Thus, apart from reliable and credible news agencies, users may be exposed to non-factual or manipulated information created by known or unknown individuals, large language processing models, tools and manipulation agents, etc., which can be hard to identify if it contains misleading or malicious information.

Below is an example of entity manipulations performed to create misinformation.

- *Human written text:*

  "Dutch intelligence warned CIA about alleged Ukrainian plot to attack Nord Stream pipelines, Netherlands' public broadcaster reports."

- *Manipulated text using GPT-2:*

  "German intelligence warned CIA about alleged Ukrainian plot to attack Nord Stream pipelines, Netherlands' public broadcaster reports."

In the example above, "Dutch" has been replaced with "German." This incorrect statement changes the context and can be used to propagate or invalidate an argument made by concerned authorities or experts.

Nowadays, news media is multi-modal, containing texts, images, videos, and soundtracks, increasing the information manipulation area. For instance, a COVID-19 anti-vaccination post can have text that talks about its developments and then attaches a graphic illustration of a dead person, which demonstrates different things and can also create an impression in the readers' minds about the vaccine being dangerous for you. Building an effective and robust method to detect inconsistencies and manipulations in multi-modal media is quite challenging as it requires evaluating each modality, cross-modal connections, and the credibility of the combinations. In some cases, although image and text may not individually be misinformative, taken together can create misinformation.

1.2 Thesis Statement

This thesis states that a two-stage forensic model can effectively identify misinformation in news media. In the first stage, we focus on image manipulation detection. The images in this stage comprise generic pictures taken by journalists and media companies that different tampering techniques have manipulated. In the second stage, we process multi-modal news articles with infographic images and related text and detect inconsistencies. We try to analyze and detect mismatches in named entities and parts of speech within the data. Due to a lack of structure within the infographics, we hypothesize using several object recognition and detection models to extract and process the data in unstructured images.

1.3 Thesis Contributions

The contribution of this thesis is two-fold, which are explained in chapters 3 and 4:

1. In Chapter 3, we work on image manipulation detection, localization, and labeling. We demonstrate various methods and how combining low-level with high-level image features effectively identifies manipulations in images. Specifically:

    - We provide an overview of image manipulation techniques and forensic clues left behind due to manipulation operations.
    - We propose a convolutional neural networks model to detect and classify manipulated images using ELA and JPEG compression analysis.
    - We then implement Nedb-Net, a noise, and edge-based manipulation localization model, to localize the regions of manipulations in the images.
    - Finally, we extend our detection module by labeling the objects using an object detection model in the localized regions to provide evidence for the manipulations.

2. In Chapter 4, we demonstrate methods that can process multi-modal data and exploit the textual and visual features to detect inconsistencies effectively in infographic news articles. Specifically:

    - We demonstrate how digital character recognition can extract data from news infographics.
    - We developed a tool to introduce inconsistencies in textual data using large language understanding models. We also introduce our novel infographic news dataset, MMA-Infographics, created using this tool.
    - We propose contradiction and entity mismatch detection algorithms that can detect inconsistent image-text pairs in news articles.

- Finally, we present our results on detecting inconsistent data in multi-modal settings without intensive feature engineering or learning-based approaches.

1.4 Outline of Thesis

This thesis is structured as follows:

1. Chapter 2 presents the background for misinformation in news media, a comprehensive and extensive collection of recent studies on manipulation and tampering techniques, features, and research advances. We also briefly present publicly available datasets for various misinformation detection tasks. We list datasets for image manipulation and image-text inconsistency detection, including a novel dataset created by us that exploits named entities and parts of speech in long/short texts and extracts textual data from infographic images to introduce controlled inconsistencies.

2. Chapter 3 presents our research on the first task of manipulation detection in images as the first stream of our forensic method. This module detects, localizes, and labels manipulations in news images. We demonstrate the use of multiple features in a deep learning model that identifies boundaries and manipulated regions.

3. Chapter 4 presents work on our algorithm's second stream that detects inconsistencies in multi-modal news media assets, majorly with infographic images. This chapter also formally defines the challenges with multi-modalities in media and the designated terminology used in this thesis.

4. Chapter 5 concludes this work and highlights directions for future work.

Chapter 2

Background and Related Work

The dissemination of manipulated or falsified information can cause chaos, hatred, and trust issues among humans and can eventually hinder the development of society. Manipulated news has negatively impacted the population, such as the 2016 US Presidential Elections, the COVID-19 pandemic, and the recent Russian attack on Ukraine. We are in urgent need of a defense mechanism against manipulated media. In this section, we introduce the fundamental theories of media manipulation and discuss more advanced patterns introduced by traditional news and social media. Specifically, we first discuss various definitions and differentiate related concepts. Next, we discuss a wide range of research and common feature representations and provide details on publicly available datasets that can be used to detect and combat manipulated news media.

2.1 Media in 21st Century

2.1.1 Media Manipulation

The concept of ”fake news” or ”manipulated news” has existed since the widespread circulation of news after the invention of the printing press in 1439. However, there is a lack of consensus on its definition. A narrow definition of manipulated news refers to intentionally and demonstrably manipulated or false articles aiming to deceive readers. This definition emphasizes two crucial aspects: authenticity and intent. Manipulated news contains verifiable incorrect information and is created deliberately to mislead consumers. Understanding the underlying intent of manipulations in the news contributes to a more comprehensive topic analysis. It eliminates ambiguities between related concepts like rumors, conspiracy theories, propaganda, and hoaxes.

2.1.2 Manipulations in traditional media

With the evolution of media platforms, from newsprint to radio and television, and more recently, the emergence of online news, the landscape of fake news and manipulated media has undergone significant changes. Humans have inherent limitations when differentiating between new and manipulated information. Various psychological and cognitive theories help explain this phenomenon and the influential power of managed news, particularly in traditional media. They primarily target consumers by exploiting their vulnerabilities. Consumers opt for ”socially safe” options when consuming and disseminating news information, adhering to established norms within their communities, even if the news being shared is false or fake.

One crucial factor that renders consumers naturally vulnerable to false information

is the realism in news articles. Consumers tend to believe that their perception of reality is the only accurate perspective, often dismissing alternative viewpoints as uninformed, irrational, or biased. This cognitive bias creates an environment where it can easily be perceived as accurate. Another significant factor is confirmation bias, wherein consumers prefer information that confirms their beliefs and opinions. This bias leads individuals to seek and trust news sources that align with their preconceived notions. Due to these cognitive biases ingrained in human nature, manipulated news often finds acceptance among consumers. Psychological studies have shown that presenting accurate, factual information to correct false narratives can sometimes prove unhelpful and may even reinforce wrong perceptions, particularly within ideological groups. Other factors like social credibility and frequency facilitate the propagation of manipulated media. Individuals are more likely to perceive a source as credible if others perceive it as credible, especially when limited information is available to assess its truthfulness. Studies have also shown that increased exposure to an idea can generate a favorable opinion, due to which consumers tend to believe the information they encounter frequently, even if false.

2.2 Feature Representations and Detection Methods

Extracting valuable features involves identifying relevant information from both the news content and the social context. This can include factors such as the language and visuals used in the news article, sensationalist or inflammatory words or objects in pictures, the source or author's credibility, and engagement patterns on social media, such as likes, shares, comments, and user profiles. By representing these features appropriately, they can serve as valuable input for machine learning models or other algorithms designed to detect manipulated news media. The

interactions, behaviors, and characteristics of users on social media platforms provide valuable insights into the credibility of news. By incorporating this metadata as auxiliary information, we can improve the accuracy and effectiveness of detection algorithms.

Traditional news articles contain the following attributes:

- Source: Author or publisher of the news article
- Headline: Short title text that aims to catch the attention of readers and describes the main topic of the article
- Body Text: Main text that elaborates the details of the news story; there is usually a major claim that is specifically highlighted, and that shapes the angle of the publisher
- Image/Video: Part of the body content of a news article that provides visual cues to frame the story. Based on these raw content attributes, different feature representations can be built to extract discriminating characteristics of fake news.

We extract feature representations from these attributes to capture specific image patterns, linguistic cues, or other indicators to help distinguish between new and manipulated news articles. The selection and combination of these representations depend on the specific requirements of the task and the chosen machine learning or analysis techniques.

2.2.1 Linguistic

Linguistic-based features can be extracted from the text content at various levels, including characters, words, sentences, and documents. Existing research has utilized standard and domain-specific linguistic features to capture different aspects

of manipulations in the news.

Common linguistic features used in natural language processing tasks include:

- Lexical item features: These features involve character-level and word-level attributes, such as the total number of words, average characters per word, frequency of long words, and the count of unique characters.
- Syntactic features: These features focus on sentence-level attributes, such as the frequency of function words and phrases (e.g., "n-grams" and bag-of-words approaches) or using punctuation and part-of-speech tagging.

Domain-specific linguistic features can also be employed, precisely aligned with the news domain. Examples of domain-specific features include quoted words, external links, the number and length of graphs or visual elements, and other characteristics unique to news articles.

In addition to these features, specific cues in writing styles can be designed to capture deceptive patterns to differentiate manipulations. This can involve using lying-detection features or other indicators highlighting inconsistencies or manipulative techniques in the articles. By combining these linguistic features, researchers and practitioners can develop robust models and algorithms to detect and characterize manipulations in the news. Among the various traditional techniques, recurrent neural networks (RNNs) such as simple RNN, GRU, and LSTM are straightforward and effective in classifying textual news. Advanced deep learning methods, for instance, Zellers et al.[43] pre-trained a generator using the same architecture as GPT-2 [29] on a large-scale news corpus, demonstrating its effectiveness in detecting neural-generated fake news. More recently, Fung et al. [8] improved the control over the generated text in a report by conditioning the generator on knowledge elements extracted from the original news article, such as entities, relations, and events. Shu et al.[33] enhanced the factual accuracy of the

generated articles by introducing a fact retriever that sourced relevant information from external corpora.

2.2.2 Visual

Manipulated news often takes advantage of people's vulnerabilities and exploits emotions by incorporating stunning or fabricated images to provoke consumers' anger or other strong emotional responses. The identification of manipulated images relies on various user-level and hand-crafted features within a classification framework. Researchers have extracted visual and statistical features for news verification purposes. The assumption of manipulation clues plays a vital role in detecting visual manipulations. Some of the common clues include edge discontinuity, lighting differences, compression artifacts, and intrinsic camera properties. These features contribute to a more comprehensive analysis and improved image manipulation detection.

- **Edge discontinuity**: Splicing-type image manipulations leave edge discontinuities around spliced regions. These regions will show sharp transitions around the edges. A blurring operation is often followed to suppress these edge artifacts. But even after blurring the edges, they may still differ from the edges of other regions developed by the camera. These blurring patterns at the edges can be used to detect image tampering. Work done in [24] assumed these edge discontinuities can be detected using bicoherence features along the horizontal and vertical axes. They proposed to generate image residuals by removing the bicoherence part from the authentic part of the image introduced due to manipulations. They demonstrated the effectiveness of these residual features to reveal better edge features than regular images. Other feature representations to detect edges artifacts include Weber local descriptor [32], LBP [1][45], steerable pyramid transform [22], and

co-occurrence matrices [37] [25].

- **JPEG compression artifacts**: Re-saving a JPEG image after manipulation operations create a double quantization effect. Authors in [28] developed statistical tools to detect these JPEG artifacts. He et al. [12] pointed out that the original image will have this double quantization effect, whereas the manipulated region will not. They proposed to analyze these quantization effects using histograms of discrete cosine transforms of manipulated regions and classify them using SVM classifiers. These histogram features were later improved in [36]. The above idea was also adopted by Wang et al. [38]. Since authentic images contain less high-frequency information than the manipulated regions, they proposed to construct JPEG compression noise maps by subtracting the original image from a JPEG compressed image and using 1-d convolution operations to detect manipulations. Armeini et al. [3] combined the 1-d convolution with 2-d convolution in RGB channels to localize these manipulations. Despite the promising results of using JPEG compression artifacts for manipulation detection, their application is limited to JPEG files only. It will fail in cases where the images are not JPEG.

- **lighting and color differences**: The lighting distribution of manipulated regions in an image might be different from the rest of the image. Johnson and Farid [15] proposed to find these regional lighting differences in images in different directions. Differences in lighting along different directions indicate image manipulations. Peng et al. [26] work improved the difference estimation accuracy. Wu and Fang et al. [40] proposed calculating the error angle between image blocks from different regions. They demonstrated that the angle difference between a block from manipulated regions and authentic regions would be more significant than two blocks from authentic regions.

- **Camera Properties**: Cameras have manufacture-specific image processing methods which leave behind camera traces or properties. Spliced regions may have different camera properties than the authentic regions of the images. These camera traces can help detect manipulations in images. The standard camera properties are photo response non-uniformity (PRNU), camera response functions (CRF), and color filter arrays (CFA). Lin et al. [18] estimated CRF from image patches and analyzed the properties to detect manipulations. The CFA unit is used to interpolate colors using demosaicing algorithms by the camera sensors. Authors in [27] used EM algorithms to find the parameters of the demosaicing algorithms to estimate the interpolation kernels. Using the estimated kernels, they detected manipulated regions in authentic images. PRNU artifacts are noise that is introduced in images during the imaging process. PRNU for images taken by different cameras will be different. Lukas et al. [19] and Chen et al. [**160**] calculated correlations between the PRNU of each image block to detect and localize manipulated regions. Some works use CNN models to detect differences in PRNU values from unknown camera models.

2.2.3 Multimodal

Real-world news is often composed of multiple modalities, like the image or a video with associated text and metadata, where information about an event is incompletely captured by each modality separately. Such multimedia data packages are prone to manipulations, where a subset of these modalities can be modified to misrepresent or re-purpose the information. Several recent models have explored the importance of multi-modal information in detecting manipulations [4]. For example, Jin et al. note jinetal focused on extracting and combining multi-modal and social context features using an attention mechanism. EANN [39] employed

an adversarial approach to learning post representations by leveraging textual and visual information, specifically removing event-specific features to improve the handling of new events. Khattar et al. [16] proposed a multi-modal variational autoencoder for rumor detection, incorporating textual and visual data. Zhang et al. [44] designed a multi-modal multi-task learning framework that included the stance task. Jaiswal et al.[14] first formally defined the multimedia semantic integrity assessment problem and combined deep multi-modal representation learning with outlier detection methods to assess whether a caption was consistent with the image in its package.

Reuben et al. [35] exploits the co-occurrences of named entities in the texts to detect possible inconsistencies in news articles. They use a visual-semantic representation of news articles to classify them as consistent. Their work assumes that the named entities in captions of news articles will also be present somewhere in the body of articles. Knowledge graph-based approaches utilize external sources, a reference dataset of unmanipulated packages as a source of world knowledge to help verify the semantic integrity of the multimedia news. Fung et al. [**fung**] demonstrated a novel method for detecting inconsistency using cross-media information consistency checking and adversarial fake information generation by knowledge graph manipulation.

2.3 Datasets

2.3.1 Datasets for Image Manipulation Detection

While the type of image manipulations may vary from splicing to duplication to removal, few standard datasets types of tampered images are widely used in news tampering detection. A few of them are listed below:

1. **Columbia**: This dataset comprises 183 original color images and 180 cut-paste images. The original images were taken using four digital cameras, and the tampered images were generated from the originals using Adobe Photoshop. All color images are stored in uncompressed TIFF format, with sizes ranging from 757x568 to 1152x768. Unlike the first dataset, these images depict complete indoor or outdoor scenes rather than just photo blocks. Tampering was still conducted without post-processing. To help identify tampered regions, edge masks are provided for each image, outlining the boundaries of the manipulated areas. Although both Columbia datasets were labeled as "image splicing," they focus on cut-paste techniques that combine two or more source images. Besides this, Columbia also released a greyscale dataset without a tampering mask, hence unsuitable for the tampering localization task.

2. **CASIA v1.0 and CASIA v2.0** [5]: The CASIA datasets are large datasets for forgery classification. They are among the first to include two kinds of manipulations in one dataset. This dataset incorporates both copy-move and cut-paste tampering techniques. Furthermore, the second dataset, CASIA v2.0, features post-processing applied to the tampered images, enhancing the tampering effect. The CASIA v1.0 dataset comprises 1721 color images with a fixed size of 384x256. Among them, 800 images are original, and 921 are tampered with. All images are saved in JPEG format. In contrast, images in the CASIA v2.0 dataset have varying sizes ranging from 240x160 to 900x600 and are available in two different file formats: TIFF for uncompressed images and JPEG for compressed images. This dataset comprises 12,614 color images, 7491 authentic images, and 5123 tampered images. Initially, the two datasets did not provide masks for the tampered images.

3. **MFC Datasets** [11]: The MFC dataset is a set of datasets developed for the Media Forensic Challenge (MFC). The MFC dataset is used to evaluate the performance of automated image and video manipulation detection and localization technologies. The MFC dataset contains over 176,000 high-provenance (HP) images.

4. **DEFACTO** [10]: This dataset is for image and face manipulation detection and localization. The dataset was automatically generated using Microsoft's everyday object-in-context database (MSCOCO) to produce semantically meaningful forgeries. Four categories of copies have been developed. They are splicing forgeries which consist of inserting an external element into an image; copy-move forgeries, where a part within an image is duplicated; object removal forgeries, where objects are removed from prints and lastly, morphing, where two images are warped and blended. Over 200000 images have been generated, and each image is accompanied by several annotations allowing precise localization of the forgery and information about the tampering process.

5. **COVERAGE** [47]: The dataset comprises 100 original images and their corresponding tampered versions. All images are stored in the TIFF format, and ground truth masks for the tampered images are provided.

6. **Wild Web Dataset** [42]: The Wild Web tampered image dataset [60] gathers images directly from the web. The tampered images in this dataset are generally more challenging for tampering localization due to additional post-processing operations, such as re-saving and resampling, that occur when the images circulate online. Although the dataset is not publicly available, researchers can access it by requesting author permission. The dataset comprises over 13,000 images with approximately 90 tampered cases, primarily

in JPEG format, with some images in PNG, GIF, or TIFF formats. All the collected images feature confirmed tampering types, with the majority being cut-paste images and a few being copy-move and erase-fill images. For each tampering case, ground truth masks were manually created.

2.3.2 Datasets for Image Text Inconsistency Detection

Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites. However, verifying the accuracy of communication is a difficult task that typically requires experts in the relevant field to analyze claims meticulously, supporting evidence, contextual information, and reports from authoritative sources. Naturally, news data with annotations can be obtained through different methods: utilizing expert journalists, relying on fact-checking websites, employing industry detectors, or engaging crowd-sourced workers. Nevertheless, there is no consensus on standardized datasets for detecting manipulated news. Some publicly available datasets are mentioned below:

1. **LIAR16** [44]: Obtained through the PolitiFact fact-checking website's API, this dataset comprises 12,836 short statements manually labeled by humans. These statements were sampled from various sources such as news releases, TV or radio interviews, campaign speeches, etc. The truthfulness labels for news range across multiple classes, providing fine-grained distinctions: ”pants-fire,” ”false,” ”barely-true,” ”half-true,” ”mostly true,” and ”true.”

2. **BuzzFeedNews15** [34]: This dataset encompasses a comprehensive selection of news articles shared on Facebook during a week surrounding the 2016 US election, specifically from September 19 to 23 and September 26 to 27. Each post and its linked article underwent fact-checking conducted by five

BuzzFeed journalists, focusing on individual claims. This dataset has been further enhanced by incorporating related articles, associated media, and relevant metadata. It comprises 1,627 pieces, including 826 mainstream articles, 356 left-wing articles, and 545 right-wing articles.

3. **BreakingNews** [30]: The BreakingNews dataset consists of approximately 100,000 articles published between January 1 and the 31st of December 2014. All reports include at least one image and cover various topics, including sports, politics, arts, healthcare, or local news. The main text of the articles was downloaded using the IJS newsfeed (Trampuš and Novak, 2012), which provides a clean stream of semantically enriched news articles in multiple languages from a pool of RSS feeds.

4. **NeuralNews** [35]: The NeuralNews dataset consists of human-generated and machine-generated articles. We build NeuralNews on top of the GoodNews dataset extracted from the New York Times to obtain human-generated articles. The Grover model generates articles using human-generated titles and articles as context. The dataset is divided into real articles with real captions, real articles with generated captions, generated articles with generated captions, and generated articles with real captions. The dataset has about 32K samples of each article type resulting in about 128K total samples.

5. **Pheme** [17]: This dataset consists of tweets collected during various breaking news events and discussions on Twitter. The dataset includes tweets related to events such as natural disasters, political controversies, and public emergencies. Each tweet in the dataset is labeled with information about its credibility, such as whether it contains accurate information, false information, a rumor, or is unverified. The Pheme dataset has been instrumental in understanding how rumors spread and how incorrect information can be

identified and countered in online social networks.

6. **SD dataset**: This is a fake news dataset focusing on the news being shared on Twitter. It consists of news article links and human judgment labels denoting if they are fake or not, as well as engaged tweets, the stance of such tweets, the publisher of the news article, and article citations by other news outlets on Twitter.

7. **InfographicVQA** [21] : This is a large infographic dataset, including 5.4k images and 30k question-answer pairs, written by humans. Images in the dataset were sourced from the internet. This dataset comprises a diverse collection of infographics and question-answer focusing on elementary reasoning and basic arithmetic skills.

8. **NewsCLIPpings** [20]: Automatically generated out-of-context image-caption pairs in news media. The images and texts are unmanipulated but mismatched. This dataset is based on VisualNews [9] and introduces Image-caption and caption-caption types of inconsistencies.

9. **MMA infographics**: Our novel dataset contains 11000 infographic image text pairs. The text includes a headline, body, date, tags, and other infographic image metadata. At the same time, the images are solely comprised of infographics like plots, charts, diagrams, flowcharts, and descriptive text. We performed controlled manipulations with the dataset to create inconsistent news image-headline pairs.

Chapter 3

Image Manipulation Detection

3.1 Overview

Depending on the content of the target images in news media, the manipulations could be divided into splicing (Figure 3.1), copy-move (Figure 3.2), or removal (Figure 3.3). One source image region is copied and pasted into other target images in image splicing. In contrast, in the copy-move and removal type of manipulations, regions of the image are either duplicated or removed from within the same image. Removal is often followed by in-painting or filling operations to cover the missing regions. These images are further resized, compressed, and enhanced, making detecting manipulation challenging for the naked eye. We will detect manipulated images and localize and label the object in manipulated regions. This will help analysts further understand the tactic and intent (e.g., detecting a weapon inserted in the image to make the scene appear more violent) of manipulation and its relation to broader campaigns.

We first detect whether the images are manipulated by using a deep convolutional neural network model by exploiting artifacts generated by performing JPEG

Figure 3.1: Image splicing type manipulation [2]

Figure 3.2: Copy-Move type Manipulation [23]

Figure 3.3: Object removal type manipulation [23]

Compression on images. Then we localize the manipulated areas using an edge-based deep learning model, which produces refined boundaries of the manipulated regions. And lastly, we label the objects within the localized area. We use publicly available datasets to train and evaluate our methods. Figure 3.4 shows the overview of our forensic method with detection, localization, and labeling nodes.

3.2 Approach

3.2.1 Detection

The first critical task of our forensic method is to detect if images are manipulated or not. We do this by exploiting the JPEG compression errors generated in manipulated images. JPEG stands for "Joint Photographic Experts Group, " one of today's most popular digital image formats. JPEG is a lossy compression format, which means that the degree of compression can be adjusted to allow a trade-off between storage size and image quality. Thus we can maintain a reasonable image quality with a massive reduction in file size. JPEG compression leaves forensic traces that can be used to determine the origin and authenticity of an image. It also introduces other compression artifacts, such as blocking artifacts, ringing effects, and blurring.

We use the approach in [7] to analyze the differences between images saved at different compression levels to check for pixel-level inconsistencies, which can indicate image manipulation. The difference is directly calculated from pixel values as follows:

$$d(x, y, q) = \frac{1}{3} \sum_{i=1}^{3} \left[f(x, y, i) - f_q(x, y, i) \right]^2 \tag{3.1}$$

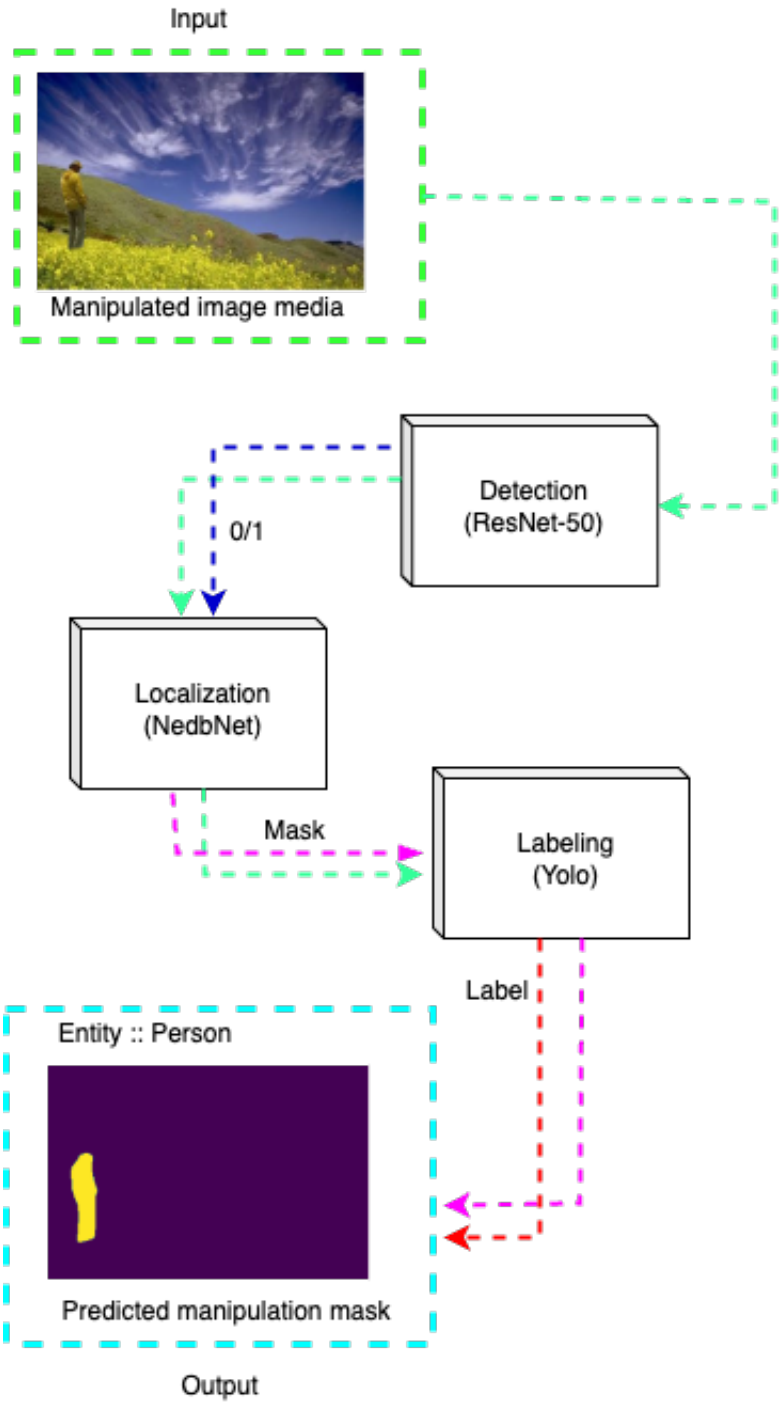
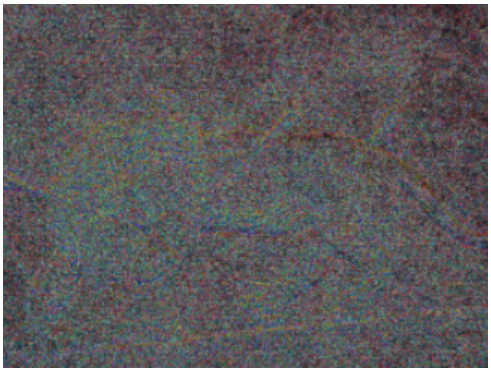
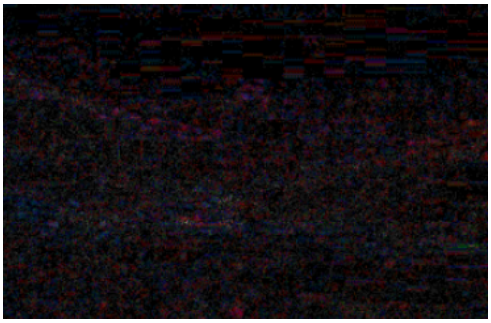


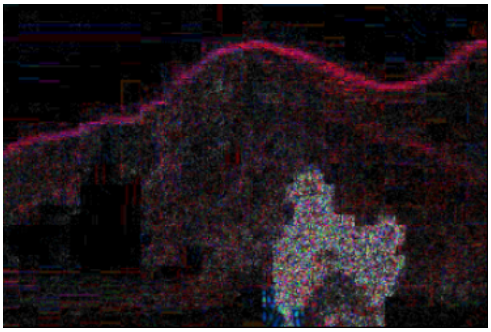
Figure 3.4: Proposed forensic method to detect image manipulations



(a) Pristine image



(b) Pristine image



(c) Manipulated image

(d) Manipulated image

Figure 3.5: ELA results on pristine and manipulated images from CASIA v2. The left column shows sample images, and the right shows their corresponding ELA output.

where $f(x, y, i), i$ = 1,2,3 denotes each of three RBG color channels, and $f_q(\cdot)$ is the result of compressing $f(\cdot)$ at quality q.

The difference would be minimum in all parts of the pristine image when saved the same number of times. However, if the image has manipulated regions, it will show variations in error levels. In Figure 3.5, we can see high activations in some regions due to differences in the error levels. a) and b) are pristine images and show minimal pixel differences, as seen on the ELA images on the right column, whereas in c) and d), we see much more pixel differences occurring due to multiple JPEG compression of manipulated images. We use this difference image with pixel-level activation as input to our CNN model, classifying them as pristine or manipulated. The images tagged as manipulated are sent to the localization module.

3.2.2 Localization

The localization module detects precise manipulation boundaries in images tagged as manipulated. The localization module uses NedbNet [46], a noise and edge-based dual-branch image manipulation detection network that uses a dual-branch network to detect subtle traces of manipulation artifacts using a high-resolution branch and a context branch (Figure 3.6). The original image from our dataset, tagged as manipulated, is first processed using an improved constrained convolution that produces a noise image. This noise image is then fed to the CNN model with a ResNet-34 backbone for localization.

One branch is used to obtain context information from images. At the same time, the other branch is used to maintain image resolution to avoid losing too many details due to the convolution operations in CNNs. Since manipulation edges are critical information for manipulation detection tasks, NedbNet uses an EEB (edge

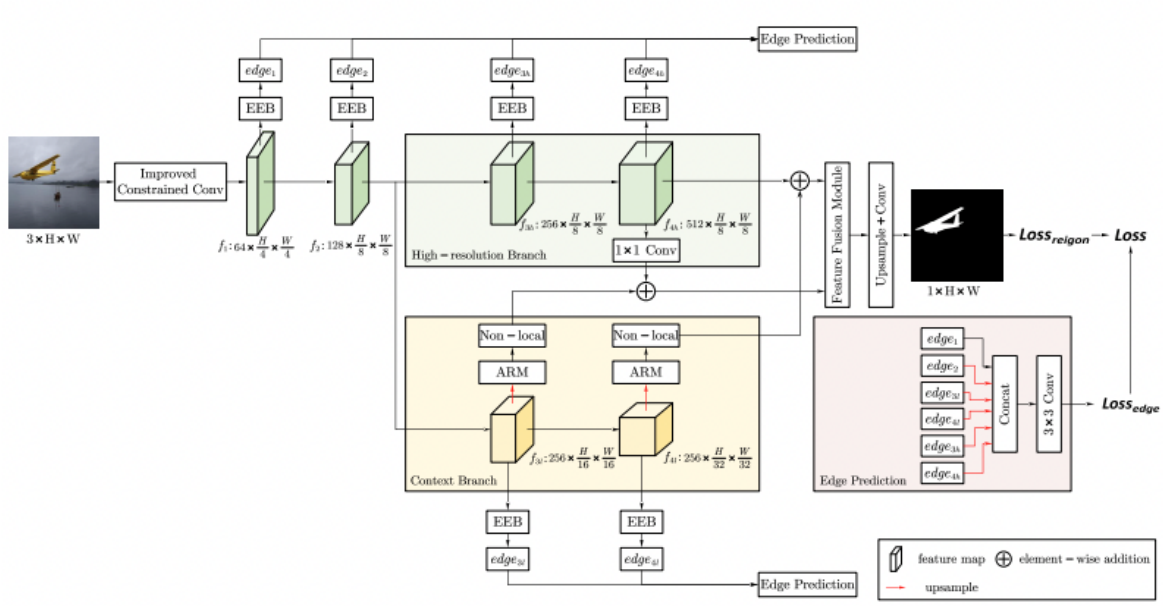


Figure 3.6: Overall architecture of NEDB-Net [46]

extraction block) 3.8 to detect manipulation edges from the features of the dual branch. The EEB extracts edges from each layer of the dual branch network and is later fused to get the final edge prediction. Conversely, the context branch uses NLM (nonlocal module) **??** that behaves similarly to self-attention to capture global correlations between pixels, essential for manipulation localization tasks.

The features from the two networks of the model, the context branch and the high-resolution branch, have different levels of information. NedbNet uses the attention refinement module (ARM) and feature fusion (FFM) inspired by BiSeNet[41] to predict the manipulation mask. The ARM is used to optimize the features in the channel dimension of the context branch, followed by a nonlocal module to perform spatial self-attention. Finally, an FFM combines the elements from both context and high-resolution branches to predict the final manipulation mask.

Improved Constrained Convolution

Generally, SRM filters [Figure 3.7] generate noise images for manipulation detection tasks. SRM filters are a set of high-pass filters which are fixed and cannot be learned. Constrained convolutions put certain constraints during weight updates, which make them behave as SRM or high-pass filters. The constraints are as follows:

$$\begin{cases} w_k(c,c) = -1 \\ \sum_{m,n \neq c} w_k(m,n) = 1 \end{cases} \tag{3.2}$$

Where w_k represents the k-th convolution kernel, (c, c) is the center position of wk, and (m, n) is the noncenter position coordinate [46]. We put constraints on the weights by calculating the sum of importance in the noncentral places and then dividing the non-center consequences by this sum. Finally, we set the center position to -1. Although this is learnable, the results from the actual training are unstable. It is unstable because the sum of the noncentral position can be harmful, and the division operation can make the positive weights harmful, which changes the input to the model layers too much. In their improved version, the kernel weights with Laplacian-like weights, but the center position is set to -1, and the other non-center positions are equal to 1 divided by the number of non-center positions. They also use the sum of absolute values of non-center positions to avoid results being negative or very small results.

Kernel size significantly influences the amount of information that is extracted. If the kernel is too large, it will capture irrelevant information and slow the computations; if it is too small, it will fail to capture important details. It is crucial to experiment over time to find the appropriate kernel size.

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 3.7: Predefined SRM Filters

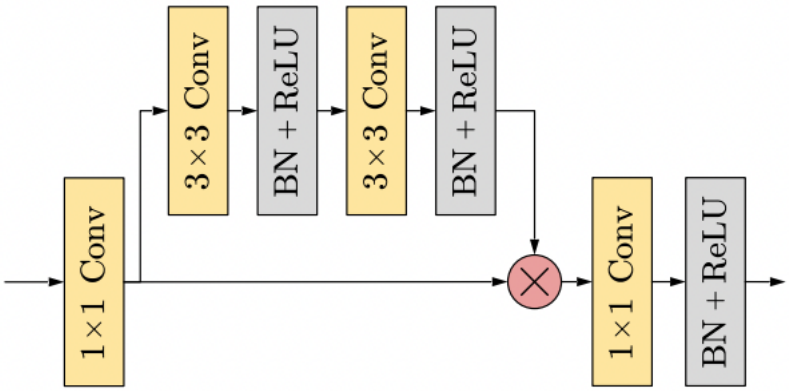


Figure 3.8: Edge extraction block [46]

Edge Extraction Block

Each layer of the backbone architecture learns different contents from the input. The EEB extracts edges from the feature outputs from each model layer. A 1x1 convolution is first used to reduce features in the depth channel, followed by a residual connection with Conv-ReLU-Conv-ReLU layers. The output is then passed through another 1x1 Conv layer, and finally, batch normalization and ReLU are applied to reduce the number of channels to 1. The below Figure 3.8 shows the EEB architecture.

Non Local Module

Pixels in images are closely related by distance. The closer two pixels are in space, the stronger their correlation is. A non-local module is used in computer vision ap-

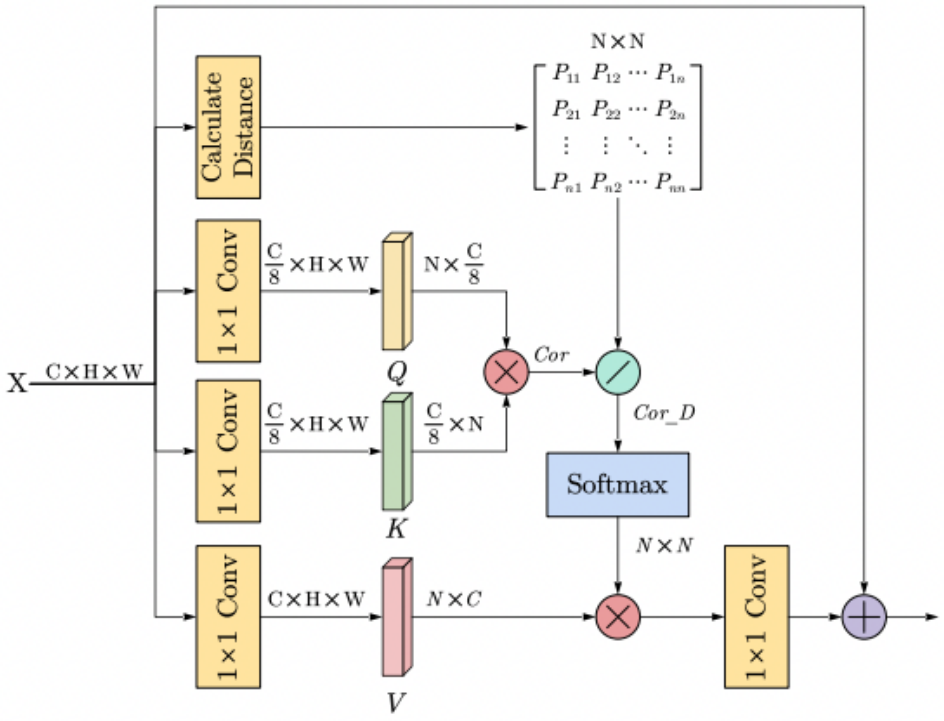


Figure 3.9: Non-local module [46]

plications to break this distance constraint and capture global correlations between pixels. The non-local module architecture can be seen in Figure 3.9

3.2.3 Labeling

After the localization module predicts the manipulation masks, we label the object in the predicted mask. We use general category objects in news media to classify the manipulated objects. Since our research problem is related to news media, we restricted our labeling module to detect only recent and common news topics. Below is the top-level taxonomy with subclasses used in model training:

- Fire/Explosion (explosion)
- Firearm/Weapon (gun, canon)
- People/Group (group)

- Person (civilian, Putin, Xi Jinping, Zelensky)
- Sign/Written Messages (signs, banner, graffiti, gesture)
- Symbol (logo, emblem)
- Vehicle (aircraft, car, motorcycle, ship)

We collect data from open-source resources for the above categories and manually annotate them using labelImg in YOLO [31] format. We train a custom Yolo-based object detection model on this dataset. Currently, YOLOv8 is the latest iteration of the YOLO family of models. YOLO stands for You Only Look Once, and these models are thus named because of their ability to predict every object present in an image with one forward pass. YOLOv8 was trained as a regression problem instead of a classification to predict the bounding box coordinates. YOLO models are pre-trained on massive datasets such as COCO and ImageNet. They provide highly accurate predictions on classes they are pre-trained in and can also learn new classes comparatively quickly. YOLO models are also faster to train and can produce high accuracy with smaller model sizes. They can be trained on single GPUs, making them more accessible to developers like us. The labeling module finally predicts object labels in the given manipulated masks.

3.3 Experimental Protocol

3.3.1 Dataset

Our detection and localization modules use CASIA v2 [5]. CASIA v2 contains high resolution 4795 images, 1701 authentic and 3274 manipulated. Figure 3.11 shows two sample images with their ground truth manipulation mask from CASIA v2 dataset. For the detection task, the labels are encoded with 0 and 1, 0 represent-

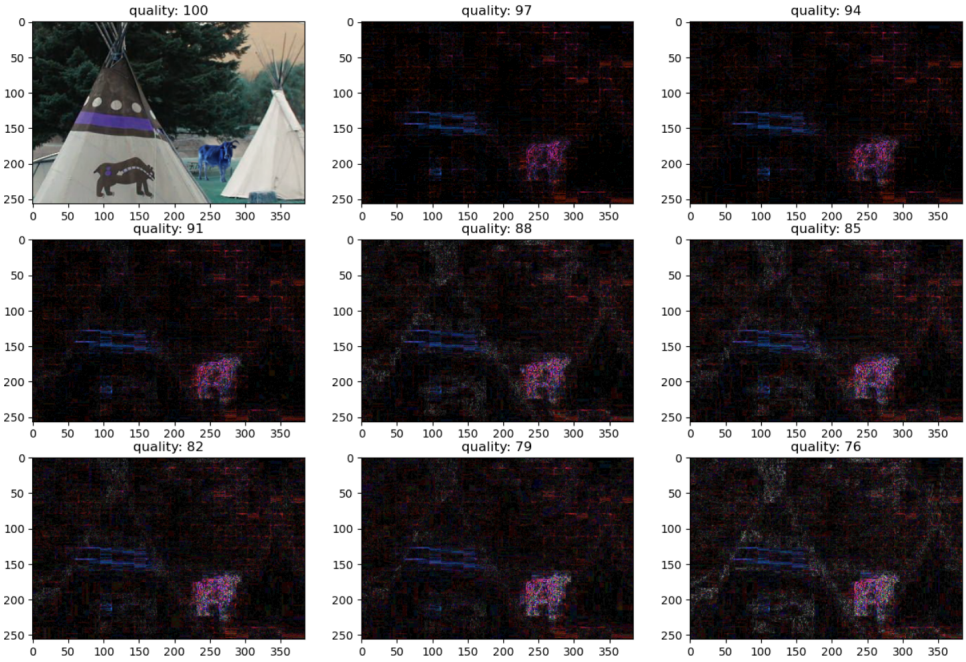


Figure 3.10: Error level analysis on a manipulated image at different JPEG compression levels.

ing pristine and one manipulated image. Then the dataset is divided into training and validation data using 80% samples for training and 20% for validation. We perform ELA on the training data and then reduce the image dimensions to 128x128. Images at a compression quality of 85 give us the best error activations. Variations in error activations at different compression quality levels can be seen in Figure 3.10. After resizing, the images are normalized by dividing each RGB value by 255.0. We use the original manipulated images normalized and resized to 512x512 dimensions for the localization task. The improved constrained convolution processes these images to obtain the noise image, which is then sent to the two branches for manipulation mask prediction.

For the labeling module, we collect and compile a new dataset. This dataset contains images of politicians, groups of people, soldiers, weapons, vehicles, signs or logos, banners, and written warnings. We scraped the data from the internet using DuckDuckGo image search API. The images were annotated using the LableImg

Figure 3.11: Sample manipulated images with corresponding ground truths, from CASIA v2.0 dataset

tool with YOLO style annotations.

3.3.2 Experimental setup

We propose a CNN model to classify these difference images as manipulated or pristine. Convolutional neural networks (CNNs) are feed-forward networks used for image-related tasks. They are handy for finding image patterns to recognize objects, classes, and categories. CNNs comprise multiple layers, including convolutional, pooling, and fully connected layers. They use a mathematical operation called convolution in place of general matrix multiplication. Here we use a Resnet50 architecture as our backbone for the classification task and train the model for 50 epochs with RMSProp optimizer and batch size 32. The critical innovation in ResNet is the concept of residual connections or skip connections. These connections enable the network to bypass specific layers and directly propagate

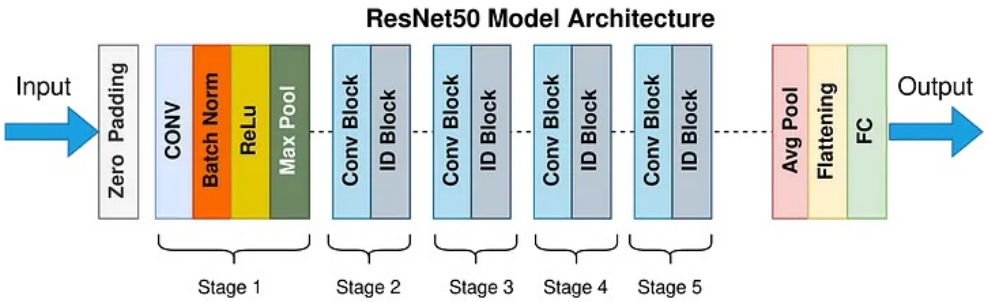


Figure 3.12: ResNet50 architecture

information from earlier layers to the last layers, allowing for the flow of gradient signals during training and addressing the vanishing gradient problem. The architecture of ResNet50 consists of a series of convolutional layers with small filters, followed by a global average pooling layer and a fully connected layer for classification. The convolutional layers are grouped into several blocks, where each block contains multiple convolutional layers and ends with a residual connection that adds the input of the block to the output. This way, the network can learn residual mappings, focusing on understanding the incremental changes to the input image data.

We optimize the model using RMSProp, a gradient-based optimization technique in training neural networks. RMSProp uses momentum and an adaptive learning rate algorithm that prevents learning rate decay too quickly. RMSProp uses exponential decay, which helps accumulated gradients focus on recent gradients instead of all previous gradients. Since the difference image generated by ELA highlights the part of the image with higher error levels and tends to have a similar color or even contrast with nearby pixels, training the CNN model becomes more efficient.

Next, The localization module is trained on the CASIA v2 dataset and is tested on CONVERGE, COLUMBIA, and NIST datasets. The high-resolution branch and the context branch use a ResNet-34 pretrianed on ImageNet. The overall loss is cal-

culated by combining manipulation region prediction loss and manipulation edge prediction loss. Dice loss is used for the region prediction to solve the problem of unbalanced pixels in manipulated and non-manipulated regions. The following calculation is used:

$$Loss = \alpha \times loss_{region} + (1 - \alpha) \times loss_{edges} \tag{3.3}$$

where the $loss_{region}$ represents the loss of the manipulation regions, the $loss_{edge}$ represents the loss of the manipulation edges, and α is the weight.

The labeling module is a Yolov8 object detection model. We divide our custom dataset into train and validation splits with an 80-20 ratio and train the model for 100 epochs. The images are resized to 640x640 and trained in batches of 16. We use SGD to optimize the training with a weight decay of 0.0005. Using a small batch significantly reduces the computational cost per iteration compared to traditional Gradient Descent methods. SGD also requires less memory to store the cost function gradients. All the models in this forensic module were trained on University at Buffalo's Deepbull servers on Nvidia RTX A5000 GPUs.

3.4 Results and Analysis

3.4.1 Detection

We train the model for 30 epochs and achieve a 79.3% classification accuracy on CASIA v2 images. We found 0.003 as the best learning rate for our training. Table 3.1 shows accuracy variations due to changes in batch size and learning rate. We train the model for 50 epochs for each experiment.

The utilization of ELA (difference) images as an image feature, along with the nor-

Model	Learning rate	Batch size	Best accuracy in %
ResNet-50	0.001	32	78.2
ResnNet-50	**0.003**	**32**	**79.3**
ResNet-50	0.005	32	78.8
ResNet-50	0.001	64	76.6
ResNet-50	0.003	64	76.8
ResNet-50	0.005	64	77.0
ResNet-18	0.001	32	74.1
ResNet-18	0.003	32	74.7
ResNet-18	0.005	32	74.5

Table 3.1: Influence on detection accuracy by using different learning rates and batch sizes

malization of RGB values for each pixel, significantly enhances the training efficiency of the CNN model. This leads to faster convergence, requiring fewer training epochs to reach a satisfactory level of convergence with high accuracy. This is evidence that the different image features effectively distinguish between pristine and manipulated images.

3.4.2 Localization

We can see refined manipulation regions in the images[Figure 3.13] predicted by NedbNet. We can see the edge detection sub-task is very helpful in capturing precise boundaries and dramatically influences the overall manipulation region localization. The global correlations maintained by the NLM self-attention module prove beneficial in segregating manipulated and un-manipulated image pixels. Since the manipulated images are compressed multiple times, some edges get too blurry to be detected by the model. This can sometimes produce broken masks with multiple small predictions. Figure 3.14 shows broken regions in the final predictions. Many studies use post-processing on the final predictions to get a connected mask using all nearby small regions.

,

Figure 3.13: Predicted manipulation masks from the localization module

3.4.3 Labeling

The trained model could detect the manipulated objects with very few false positives. The per-class confusion matrix and the precision-recall curves can be seen in Figure 3.15 and Figure 3.17, respectively. The model achieved approximately 0.78 precision and 0.68 recall Figure 3.16. Since some classes like People/Group and Person were ambiguous, some misdetections existed, along with a few cases where the model confused between signs, banners, written warnings, and some small objects due to low image resolution. These issues could be handled using a larger, high-resolution object dataset.

Figure 3.14: Broken manipulation masks predicted from the model

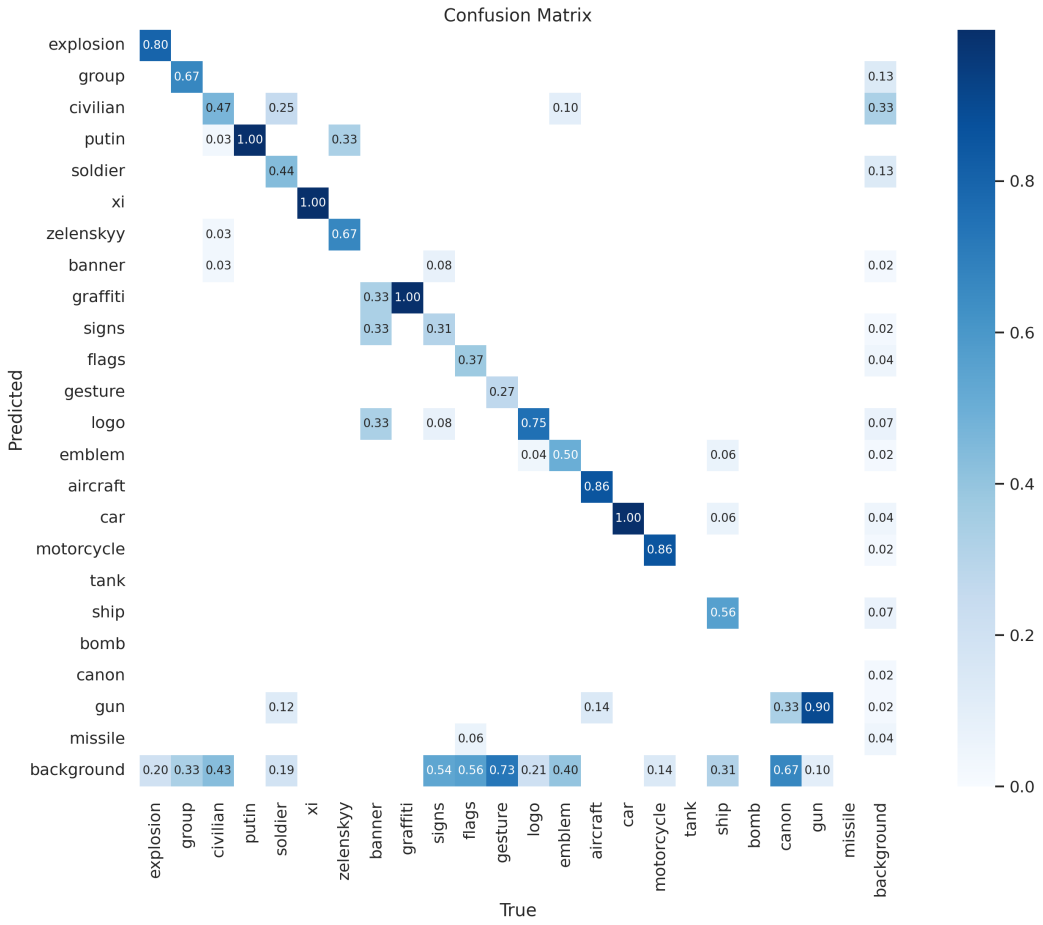


Figure 3.15: Per class Confusion matrix

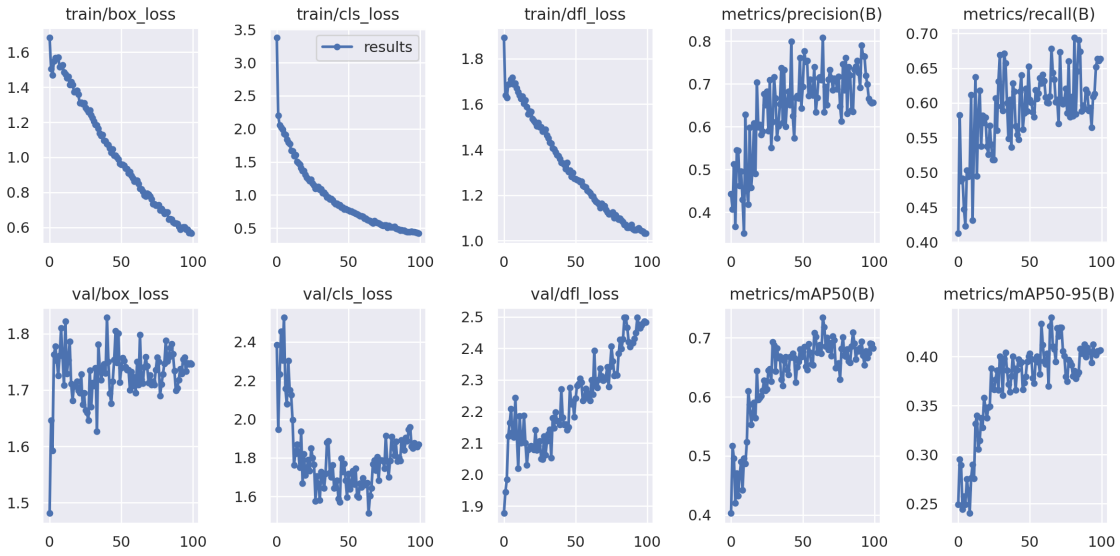


Figure 3.16: Training results from labeling model

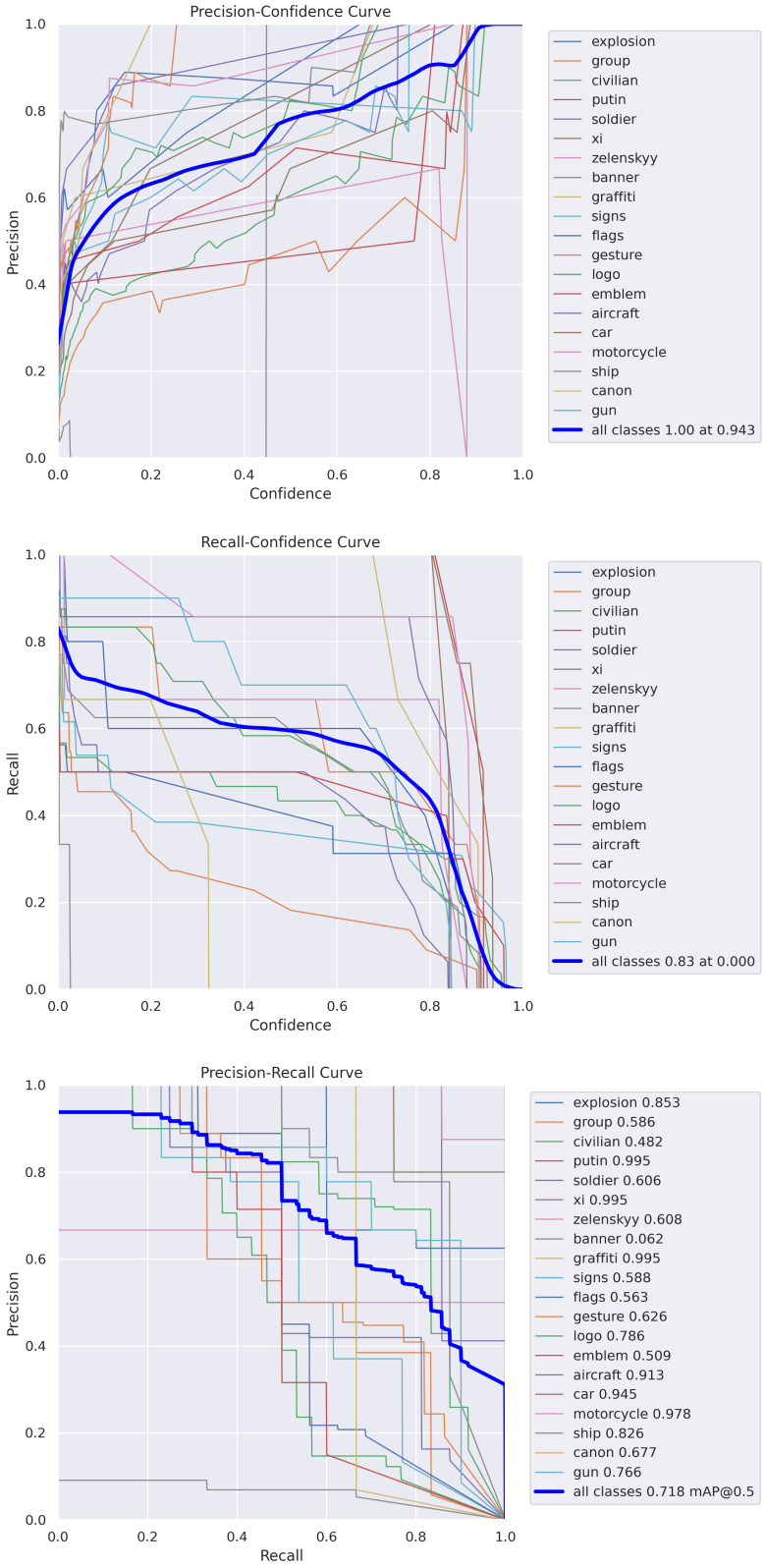


Figure 3.17: Per class PC, RC, and PR curves

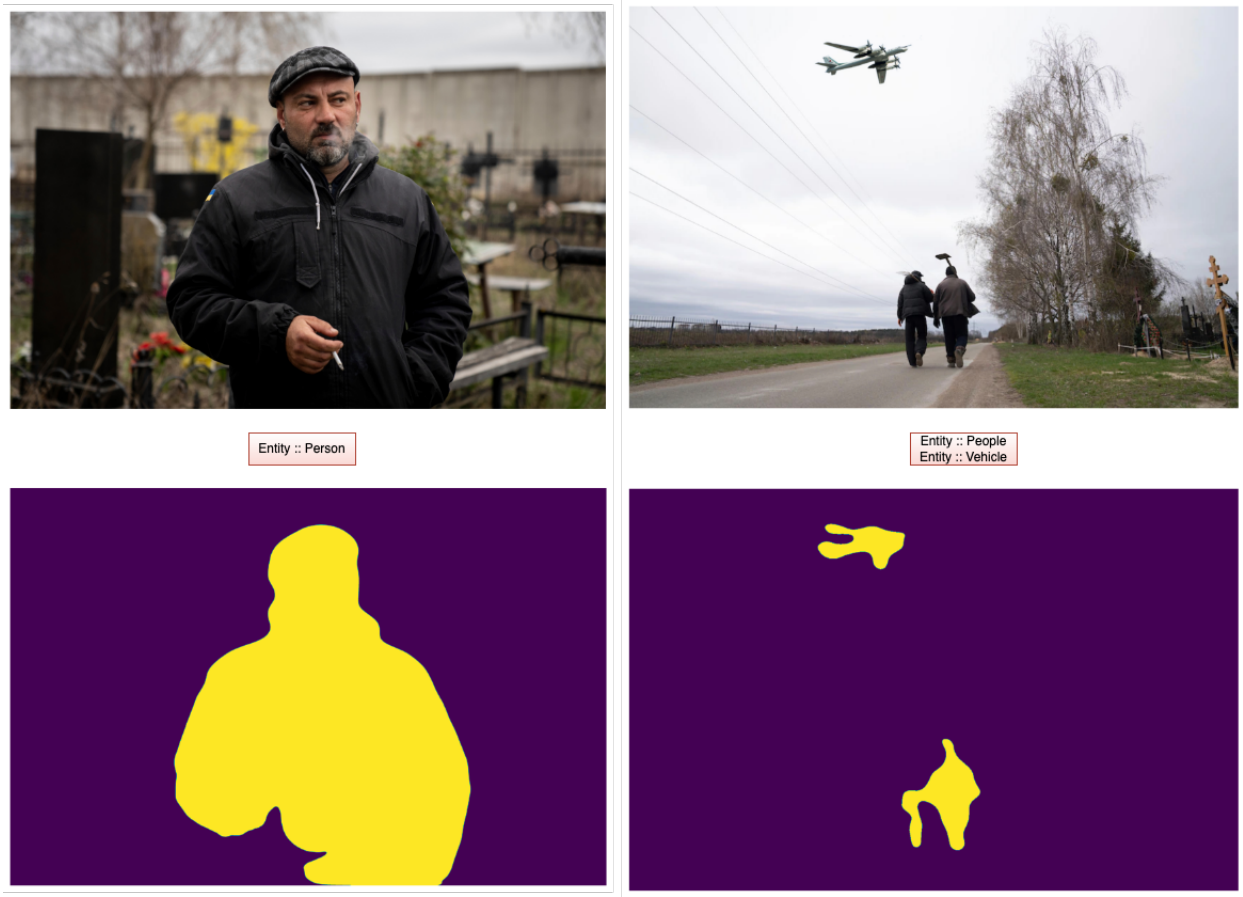


Figure 3.18: Shows final prediction mask and labeled manipulated objects

Chapter 4

Image Text Inconsistency Detection

4.1 Overview

While there are several existing approaches for inconsistency detection in news articles, they are generally limited to textual data. This chapter introduces the challenge of detecting image-text inconsistency in news articles with infographic images. Infographics present much information in a condensed format and may convey incomplete information due to oversimplification. The interpretation of data in infographics is also subjective and can vary from person to person. The visual cues can be interpreted differently, leading to varying understandings of their information. There are active researches in the field of multimodal learning which try to embed image-text pairs in the same latent space and find separation among them. These approaches are being used in social media and news domains actively. However, due to the complex representation of information in infographics, current multimodal approaches tend to be ineffective. This work present will provide a baseline approach that can be a practical reference for future inconsistency detections task with infographic images.

4.2 Approach

In this work, we try to identify inconsistencies between infographic images and their associated text in news articles. Infographic images are made up of a combination of small amounts of text and visual elements like images, symbols, and data visualizations intended to be understood easily by viewers. This makes it rather challenging due to the unstructured format, and finding any semantic relationships with the data within an image or the associated text in news articles becomes very difficult. We propose a general and relatively simple approach based on capturing data from infographic images using computer vision and natural language processing models and performing controlled comparisons between image-text to find inconsistencies in news articles.

4.2.1 Entity Mismatch

We use entity mismatch methods to detect inconsistent named entities in the texts and OCR. We first detect entities using Spacy from the selected pair of sentences. The detection types are listed in Table 4.1. We then perform comparisons of detected entities within the same entity type. For example, we compare a date-type entity with data type only. Co-reference resolution in the preprocessing stage is helpful here, which resolves context in the entire text depending on the cluster(entity) type making the comparisons more efficient. We separately evaluate numerical and alphabetical type entities to prevent false detections.

4.2.2 Contradiction

We use the contradiction detection method to detect inconsistent parts of speech words in sentences. From the selected pair of sentences from news text and OCR, we detect parts of speech keywords and perform word expansion to create antonym

and synonym sets for each extracted word and store them in a temporary cache. We tag sentences as probe and target. POS words of similar type are cross-compared from the probe and target to detect inconsistencies. If the antonym of a probe POS word is present in the synonym set of the target POS word, we mark the pair as inconsistent. This method takes two runs to complete. In the second run, we swap the tags of the sentences and rerun the algorithm. We limit the size of synonym and antonym sets to make the comparison speed efficient.

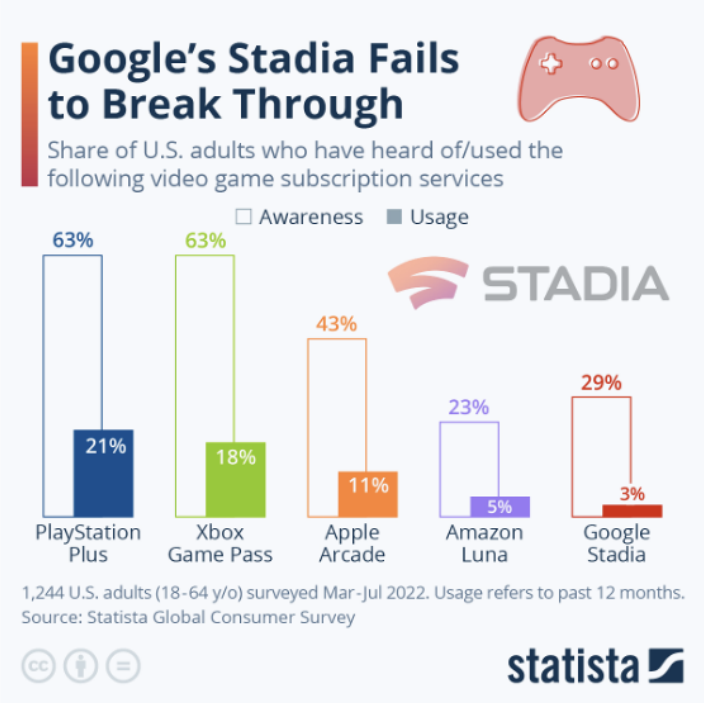
4.3 Experimental Protocol

4.3.1 Dataset

Since no public news datasets that use infographic images are available, we collected and created our dataset. We scraped articles from websites that provided statistical news. Infographics or charts usually accompany these articles to depict data on images along with textual descriptions. The raw dataset consisted of 11000 news articles. Each article had an info graph, news title, body, news type, and metadata. We introduced controlled inconsistencies in them by extracting and manipulating the named entities and parts of speech in long, short texts and OCR extracted from infographics. We also performed random text manipulations on a small dataset subset where we swapped texts from two random news articles.

4.3.2 Text in Image

Infographic images are visual representations that combine text, images, and graphics to convey information or data. They are visually appealing but unstructured and complex. We must extract this textual information from the infographic im-



Headline : Google's Stadia Fails to Break Through

Body : Less than three years after its launch in November 2019, Google is shutting down its cloud gaming service Stadia. Despite arriving just in time for the video game boom that was ignited by the Covid-19 pandemic, the service never gained traction with consumers, as cloud gaming has yet to reach its full potential. According to findings from Statista's Global Consumer Survey, Stadia trailed other gaming subscription services both in terms of recognition and usage. Among a sample of 1,244 U.S. adults surveyed between March and July 2022, just 29 percent had heard of Stadia, while a meager 3 percent had used the service in the past 12 months. That's a longshot from the more traditional PlayStation Plus and Xbox Game Pass subscriptions, which 63 percent of respondents were familiar with. While Stadia's approach to streaming games for consumers was built on a strong technology foundation, it hasn't gained the traction with users that we expected so we've made the difficult decision to begin winding down our Stadia streaming service , Google wrote in a blog post. In an applaudable gesture of gratitude towards Stadia's early adopters and users, Google will be refunding all Stadia hardware purchases made through the Google Store as well as all game and add-on purchases made through the Stadia store.

Figure 4.1: Sample News article from our MMA infographics dataset

ages to build our detection method. Extracting data from infographic images can be more challenging than extracting text from regular images since the data is often embedded within graphical elements, and maintaining order is difficult. We use Paddle OCR [6] to extract texts from the images. Paddle OCR does Layout Information Extraction and Key Information Extraction. It integrates an image direction correction module and a layout restoration module to enhance the algorithms' functionality and get better performance. The critical information extraction module performs Semantic Entity Recognition (SER) and Relation Extraction (RE) to get critical entities in images. In contrast, the layout information extraction module analyzes the layout to get title, paragraph, and table objects. We extracted and saved the OCR data in a separate JSON file.

4.3.3 Data manipulation

We perform controlled and consistent replacement in texts to create inconsistency in the collected data (both news text and OCR from image). We used Spacy [13] for this. Spacy is an open-source library for Natural Language Processing (NLP) written in Python, which provides a wide range of tools and functionalities for various NLP tasks, such as tokenization, part-of-speech tagging, named entity recognition, syntactic parsing, etc. We use it to detect named entities and parts of speech and perform controlled in-domain replacements in news texts and the infographics OCR. We create a database of pre-compiled words of in-domain entities and parts of speech from the scraped data and the internet and use it for replacements. We also perform random replacements where we randomly swap two similar-sized texts with each other. Below we briefly describe the replacement processes.

- **Named Entity Replacement**: We create a database of named entities used during the replacement process. This database is compiled using Spacy from within our dataset and from other open resources, e.g., Wikipedia. Next, we

Table 4.1: Named entity types used in text manipulation

ENTITY	Description
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage, including ”%”
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	”first,” ”second,” etc.
CARDINAL	Numerals that do not fall under another type
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
ORG	Companies, agencies, institutions, etc.
LOC	Non-GPE locations, mountain ranges, bodies of water
GPE	Countries, cities, states
PRODUCT	Objects, vehicles, foods, etc. (not services)
FAC	Buildings, airports, highways, bridges, etc.
LAW	Named documents made into laws.

Table 4.2: POS types used in text manipulation

POS	Description
ADJ	Adjective
ADV	Adverb
NOUN	NOUN
VERB	Verbs

read our data and detect available entity types in it. We randomly select an entity type and perform controlled replacements in either of the three text types: headline, body, or OCR. In the case of long texts like the body of news articles, we make sure that every instance of that target entity is consistently replaced with the same replacement entity throughout the text. We do not change any metadata of the article. The below table[table] shows the entity types that are manipulated in the texts:

- **Parts of Speech Replacement**: Similar to Named Entity Replacements, we create a separate database using our dataset as key-value pairs for parts of speech replacements. We use the four types of POS replacement in our

	Pristine	Manipulated
Headline	Europe's independent hotels are slowly displaced by chains	China's independent hotels are slowly displaced by chains
Body	There are close to 3.5 independent hotels for every chain property in Europe, but this number is slowly eroding. Considering properties with 25 rooms or more, the number of independent hotels in Europe has dropped from just over 51 thousand in 2015 to 49,500 in 2021. On the flipside, hotel chains have been gaining ground, aided by the tailwind of the travel & tourism sector becoming increasingly more international.	There are close to 3.5 independent hotels for every chain property in China, but this number is slowly eroding. Considering properties with 25 rooms or more, the number of independent hotels in China has dropped from just over 51 thousand in 2015 to 49,500 in 2021. On the flipside, hotel chains have been gaining ground, aided by the tailwind of the travel & tourism sector becoming increasingly more international
OCR	Chains vs. independent properties in the Europe, by number of properties	Chains vs. independent properties in Europe , by number of properties

Figure 4.2: Example for NER type manipulation

dataset, listed in 4.2. The keys are the detected POS words, and the values are words or phrases that negate the semantic meaning of the keys. Next, we consistently replace the POS detected in our dataset using this key-value database throughout the text. The consistency of replacements is helpful in accurately characterizing the detected inconsistencies.

- **Random Replacement**: We also perform random swapping of similar type texts, e.g., headline-to-headline and body-to-body. We read two random data in pairs and swapped the texts in them.

4.3.4 Experimental Setup

We first extract textual data from infographics. Any sentence with less than two words is discarded. Then we concatenate the headline and body of the news article to form a long text. The long text and OCR text are then sent to the co-reference module. Co-reference resolves the pronouns by linking the references of an entity

	Pristine	Manipulated
Headline	Mortgage Rates Climb to Highest Level Since 2008	Mortgage Rates Climb to lowest Level Since 2012
Body	Mortgage rates in the United States continue their steep climb, adding to the woes of would-be home buyers that are already facing historically high prices and steep competition in the tight housing market. According to Freddie Mac, the average rate for a 30-year fixed mortgage increased to 6.02 percent in the week ended September 14, the highest it's been since November 2008. Along with the Fed's aggressive rate hikes, mortgage rates have climbed by almost 3 percentage points this year, threatening to push more and more potential buyers out of the market, especially as high rents and other costs of living make it increasingly difficult to save for a significant down payment.	Mortgage rates in the United States continue their steep climb, adding to the woes of would-be home buyers that are already facing historically high prices and steep competition in the tight housing market. According to Freddie Mac, the average rate for a 30-year fixed mortgage increased to 6.02 percent in the week ended September 14, the lowest it's been since November 2012. Along with the Fed's aggressive rate hikes, mortgage rates have climbed by almost 3 percentage points this year, threatening to push more and more potential buyers out of the market, especially as high rents and other costs of living make it increasingly difficult to save for a significant down payment.
OCR	Average 30 year fixed mortgage rate in the United States climbs to highest level.	Average 30 year fixed mortgage rate in the United States climbs to highest level.

Figure 4.3: Example for POS type manipulation

and replacing them with the entities they are referring to. This helps in context understanding natural language problems.

Due to the unstructured format of infographic images, the data collected using OCR will be unstructured too. Since the flow of information between different parts of infographics is challenging to identify, we perform our detection experiments in paired sentence sets. We use a BERT-based sentence transformer to create sentence-level embeddings and use cosine similarity to pair the most similar sentences from the long and OCR text. We tokenize and create word n-grams of keywords for each pair of sentences and rank the pairs using the Jaccard similarity score. We use the NLTK library for tokenizing and keyword detection. We finally select the top sentence pair for inconsistency detection. The entire data processing process is shown in Algorithm 1. This sentence pair is sent to the mismatch and contradictions detection modules for inconsistency detection. We mark the news article as inconsistent if any of the two methods return True.

Algorithm 1 Image-Text Data Processing for Inconsistency Detection Algorithm

Input: Two lists of texts, $text1(news\ text)$ and $text2(OCR)$
Output: Boolean; $true$ = inconsistent, $false$ = consistent

foreach t ***in*** $text1$ **do**
  $t \leftarrow$ coref(t)
end
foreach t ***in*** $text2$ **do**
  $t \leftarrow$ coref(t)
end

Initialize an empty dictionary $sentence\_sim$

foreach $t1$ ***in*** $text1$ **do**
  $embed\_t1 \leftarrow$ embed$(t1)$
  foreach $t2$ ***in*** $text2$ **do**
    $embed\_t2 \leftarrow$ embed$(t2)$
    $sim\_score \leftarrow$ cosine$(embed_t1, embed_t2)$
    $sentence\_sim[sim\_score] \leftarrow (t1, t2)$
  end
end

Sort $sentence\_sim$ in descending order and store the top 5 results in $sorted\_sentence\_sim$

$top \leftarrow$ sorted_sentence_sim$[0]$ $temp\_score \leftarrow -\infty$
foreach $pair$ ***in*** $sorted\_sentence\_sim$ **do**
  $A \leftarrow$ get_keywords$(pair[0])$ $B \leftarrow$ get_keywords$(pair[1])$
  $word\_sim\_score \leftarrow$ jaccard(A, B)
  if $word\_sim\_score > temp\_score$ **then**
    $top \leftarrow pair$ $temp\_score \leftarrow word\_sim\_score$
  end
end

$sent1 \leftarrow$ top$[0]$ $sent2 \leftarrow$ top$[1]$

return *contradiction*$(sent1, sent2)$ ***or*** *mismatch*$(sent1, sent2)$

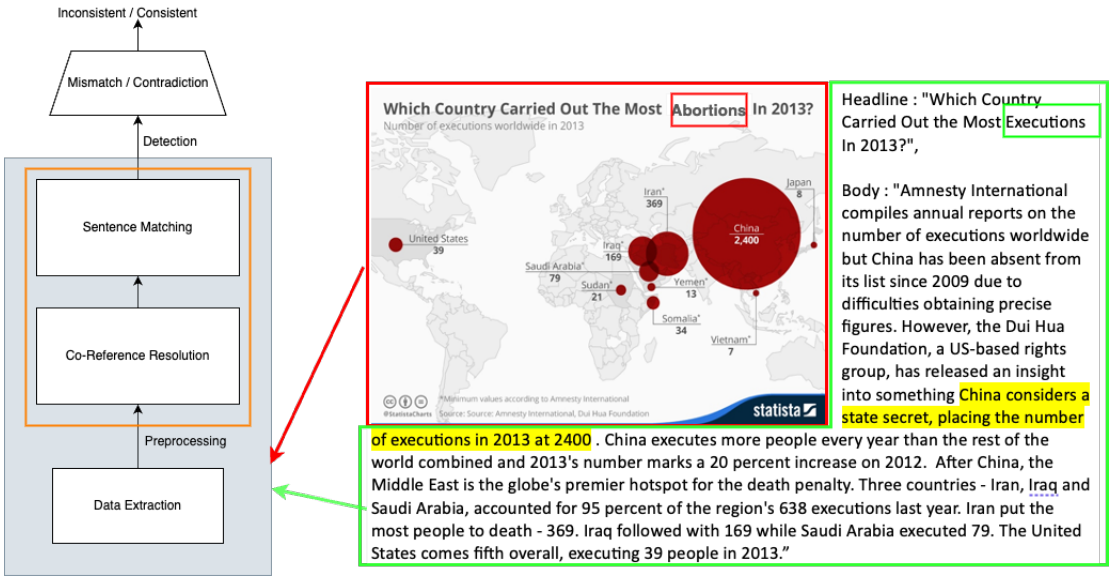


Figure 4.4: High-level design of our inconsistency detection method

4.4 Results and Analysis

We tested our inconsistency detection algorithm on the MMA infographics dataset. We combined pristine and manipulated samples to create a mixed dataset with 10000 samples, 5000 of which were pristine, and the rest were equally distributed manipulated samples. Our baseline algorithm detected pristine from manipulated samples with an average accuracy score of around 18.68%. We can see the confusion matrix (figure 4.5) and the classification scores (Table 4.3) below. While the detection accuracy is low, we can hypothesize the connection between data extraction and text preprocessing methods. PaddleOCR fails to detect white spaces between words and punctuation symbols. This breaks the coherency and flow of data and thus affects the sentence-matching and final detection results. We can build an effective image-text inconsistency detection algorithm with better data extraction and efficient post-processing algorithms.

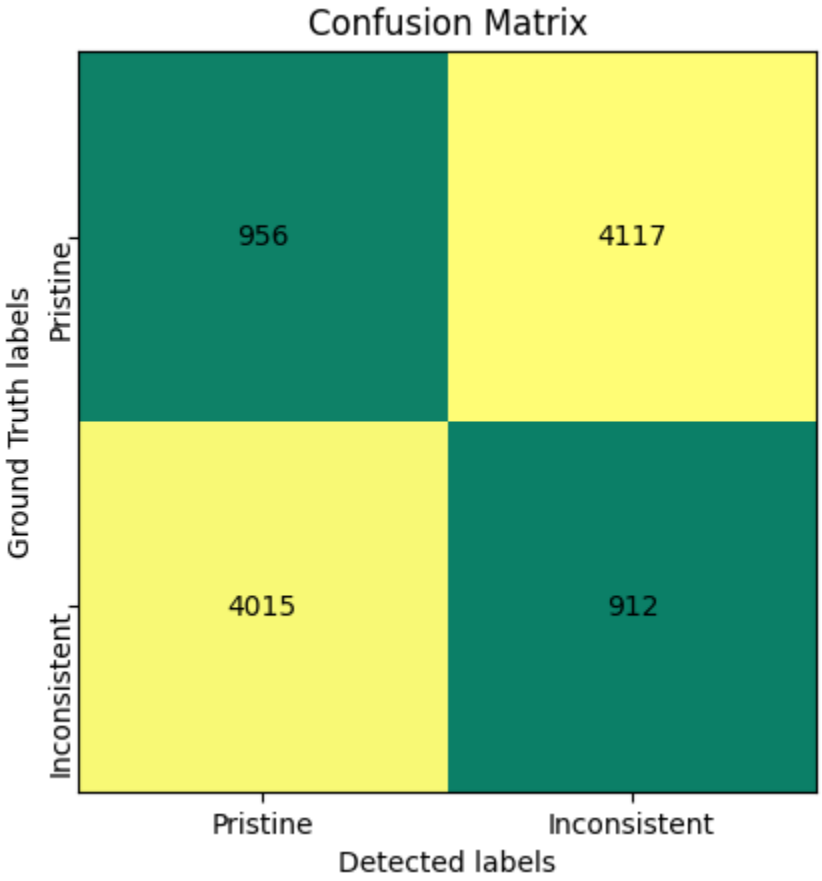


Figure 4.5: Confusion matrix for inconsistency detection with 10000 samples

	Precision	**Recall**	**F1-Score**	**Samples**
Pristine	0.19	0.19	0.19	5073
Inconsistent	0.18	0.19	0.18	4927

Table 4.3: Classification report for inconsistency detection

Chapter 5

Conclusion

From loss of trust to misleading decision-making, manipulated news can significantly impact consumers. Evaluating the data and critically ensuring information integrity is essential to mitigate the impact of manipulated news. We urgently need methods that detect manipulations with the rise of generative models and manipulation tools that can produce convincing news tamperings. Detecting tampered information is difficult, but with closer attention and recent state-of-the-art algorithms, we can identify traces or clues in the data left behind by manipulation operations. Several artifacts in manipulated data, like color consistency, pixel inconsistency and camera metadata in images, or semantic and grammatical inconsistencies in textual data, can be used effectively to detect manipulations. In this thesis, we proposed a method that detects, localizes, and labels manipulated objects in images and a baseline multi-modal inconsistency detection algorithm to detect mismatches between images and associated texts in news articles.

We first describe the image manipulation detection method. This method detects if the image is manipulated and tells us where and what the manipulation is. In the first phase, we use error-level analysis to generate low-level feature represen-

tation and then pass it to a convolutional neural network with a Resnet backbone to classify whether the images are manipulated. Following the second module, we localize the manipulated regions in images using noise and edge-based features in images. The edge extraction block and non-local module used in the NedbNet help hugely in a more precise manipulation boundary detection. The high-resolution branch aids the low-resolution output from the context branch and dramatically improves the overall manipulation prediction masks. Lastly, the labeling module provides labels to objects in manipulated regions that help in the intent and tactic behind the manipulations.

Although the method performs well, a few challenges are yet to be considered. For example, the detection module can handle only JPEG-type images, which would fail if images are formatted in other image types. On the other hand, the localization module can sometimes produce broken masks, which may be part of a larger region. Although there are several post-processing methods of handling this issue, a robust and integrated approach is yet to be researched. One future direction to this can be using conditional random fields; statistical models used in pattern recognition problems to get structured predictions.

In chapter two, we demonstrated our work on image-text inconsistency detection. We extracted textual data from unstructured infographics and performed controlled manipulation in long, short, and OCR texts to create an inconsistent dataset. To create an inconsistent dataset, we performed named entity, part of speech, and random text replacements. This novel dataset was used in our image-text inconsistency detection model. We used large language understanding models to pre-process, link, and detect inconsistent sentence pairs. We also experimented with and combined different similarity measures and thresholds to select the most similar in-context sentences in long texts. We use two different methods for detec-

tion. First, a mismatch detection method that detects the named entities; second, a contradiction detection method that works with parts of speech type manipulations.

Our detection scores show that the baseline method can detect inconsistencies in news articles; however, the algorithm is slow, and the score is low due to significant dependencies on the language understanding models and preprocessing methods. The detection also hugely depends on the data extraction from infographics. Current OCR models fail to maintain coherency and structure in the text extracted from infographics, which hugely hinders sentence similarity and hence detection results. We can make our data extraction and preprocessing processes more efficient using character recognition methods that maintain the order of text clusters in infographics and larger language models that can better link entities in sentences and maintain more extended contexts.

Reference

[1] A.A. Alahmadi et al. "Splicing image forgery detection based on DCT and local binary pattern". In: *Proceedings of GlobalSIP, IEEE, pp. 253-256* (2013).

[2] Loai Alamro and Nooraini Yusoff. "Copy-Move Forgery Detection using Integrated DWT and SURF". In: *Journal of Telecommunication, Electronic and Computer Engineering* (2017).

[3] I. Amerini et al. "Localization of jpeg double compression through multi-domain convolutional neural networks". In: *arXiv preprint arXiv:170601788.* (2017).

[4] Juan Cao et al. "Exploring the role of visual content in fake news detection. D". In: *Disinformation, Misinformation, and Fake News in Social Media,* (2020).

[5] J. Dong, W. Wang, and T. Tan. "CASIA Image Tampering Detection Evaluation Database". In: *IEEE China Summit and International Conference on Signal and Information Processing, Beijing, China* (2013).

[6] Yuning Du et al. "PP-OCR: A Practical Ultra Lightweight OCR System". In: *ArXiv* (2020).

[7] Hany Farid. "Exposing Digital Forgeries From JPEG Ghosts". In: *IEEE Transactions on Information Forensics and Security* (2009).

[8] Yi Fung et al. "InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2021).

[9] Liu Fuxiao et al. "Visualnews: A large multi-source news image dataset". In: *ArXiv* (2020).

[10] B. TAJINI G. MAHFOUDI et al. “DEFACTO: Image and Face Manipulation Dataset”. In: *27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain* (2019).

[11] H. Guan et al. “MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV 2019), Waikola* ().

[12] J. He et al. “Detecting doctored JPEG images via DCT coefficient analysis”. In: *Proceedings of ECCV, pp. 423-435* (2006).

[13] Matthew Honnibal et al. “spaCy: Industrial-strength Natural Language Processing in Python”. In: (2020).

[14] Ayush Jaiswal et al. “Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text”. In: *CoRR* (2017).

[15] E. Kee and H. Farid. “Exposing digital forgeries from 3-d lighting environments”. In: *Proceedings of WIFS, IEEE, pp. 1-6* (2010).

[16] Dhruv Khattar et al. “MVAE: multimodal variational autoencoder for fake news detection”. In: *The World Wide Web Conference, WWW* (2019).

[17] Elena Kochkina, Mariam Liakata, and Arkaitz Zubiaga. “PHEME dataset for Rumour Detection and Veracity Classification.” In: (2018).

[18] Z. Lin et al. “Detecting doctored images using camera response normality and consistency”. In: *Proceedings of CVPR, IEEE, pp. 1087-1092* (2005).

[19] J. Lukáš, J. Fridrich, and M. Goljan. “Detecting digital image forgeries using sensor pattern noise”. In: *Proceedings of the SPIE, vol. 6072, p. 15* (2006).

[20] Grace Luo, Trevor Darrell, and Anna Rohrbach. “NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media”. In: *arXiv:2104.05893* (2021).

[21] Minesh Mathew et al. “InfographicVQA”. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2021).

[22] G. Muhammad et al. “Image forgery detection using steerable pyramid transform and local binary pattern”. In: *Mach. Vis. Appl., 25 (4), pp. 985-995* (2014).

[23] Kamyar Nazeri et al. “EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning”. In: (2019).

[24] T.T. Ng and S.F. Chang. "A model for image splicing". In: *Proceedings of ICIP, vol. 2, IEEE, pp. 1169-1172* (2004).

[25] T.H. Park et al. "Image splicing detection based on inter-scale 2D joint characteristic function moments in wavelet domain". In: *EURASIP J. Image Video Process., 2016 (1), p. 30* (2016).

[26] B. Peng et al. "Optimized 3d lighting environment estimation for image forgery detection". In: *IEEE Trans. Inf. Forensics Secur., 12 (2), pp. 479-494* (2017).

[27] A.C. Popescu and H. Farid. "Exposing digital forgeries in color filter array interpolated images". In: *IEEE Trans. Signal Process., 53 (10), pp. 3948-3959* (2005).

[28] A.C. Popescu and H. Farid. "Statistical tools for digital forensics". In: *Proceedings of IH, vol. 3200, Springer, pp. 395-407* (2004).

[29] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[30] Arnau Ramisa et al. "The BreakingNews Dataset". In: *Association for Computational Linguistics* (2017).

[31] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[32] S.Q. Saleh et al. "Evaluation of image forgery detection using multi-scale weber local descriptors". In: *Proceedings of ISVC, Springer, pp. 416-424* (2013).

[33] Kai Shu et al. "Fact-Enhanced Synthetic News Generation". In: *CoRR* (2020).

[34] Kai Shu et al. "Fake News Detection on Social Media: A Data Mining Perspective". In: *Association for Computing Machinery* (2017).

[35] Reuben Tan, Bryan A. Plummer, and Kate Saenko. "Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News". In: *Empirical Methods in Natural Language Processing (EMNLP)* (2020).

[36] V.L. Thing, Y. Chen, and C. Cheh. "An improved double compression detection method for JPEG image forensics". In: *Proceedings of ISM* (2012).

[37] W. Wang, J. Dong, and T. Tan. "Effective image splicing detection based on image chroma". In: *Proceedings of ICIP, IEEE, pp. 1257-1260* (2009).

[38] W. Wang, J. Dong, and T. Tan. "Tampered region localization of digital color images based on jpeg compression noise". In: *Proceedings of IWDW, Springer, pp. 120-133* (2010).

[39] Yaqing Wang et al. "EANN: event adversarial neural networks for multi-modal fake news detection." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, KDD* (2018).

[40] X. Wu and Z. Fang. "Image splicing detection using illuminant color inconsistency". In: *Proceedings of MINES, IEEE, pp. 600-603* (2011).

[41] C. Yu et al. "BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation". In: *ECCV* (2018).

[42] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. "Detecting image splicing in the wild (WEB)". In: *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (2015).

[43] Rowan Zellers et al. "Defending against neural fake news". In: *Advances in Neural Information Processing Systems 32* (2019).

[44] Huaiwen Zhang et al. "Multi-modal knowledge aware event memory network for social media rumor detection". In: *Proceedings of the 27th ACM International Conference on Multimedia, MM, Nic* (2019).

[45] Y. Zhang et al. "Image-splicing forgery detection based on local binary patterns of DCT coefficients". In: *Secur. Commun. Netw., 8 (14), pp. 2386-2395* (2015).

[46] Zhongyuan Zhang et al. "Noise and Edge Based Dual Branch Image Manipulation Detection". In: *ArXiv* (2022).

[47] Lilei Zheng, Ying Zhang, and Vrizlynn L.L. Thing. "A survey on image tampering and its detection in real-world photos". In: *Journal of Visual Communication and Image Representation* (2019).