MULTILINGUAL, MULTIMODAL AND EXPLAINABLE APPROACHES FOR AUTOMATED FACT-CHECKING PROBLEM

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

RECEP FIRAT ÇEKINEL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JANUARY 2025

Approval of the thesis:

MULTILINGUAL, MULTIMODAL AND EXPLAINABLE APPROACHES FOR AUTOMATED FACT-CHECKING PROBLEM

submitted by **RECEP FIRAT ÇEKINEL** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Pınar Karagöz
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering, METU _____

Prof. Dr. Pınar Karagoz
Computer Engineering, METU _____

Prof. Dr. Sinan Kalkan
Computer Engineering, METU _____

Prof. Dr. Suat Özdemir
Computer Engineering, Hacettepe _____

Prof. Dr. Fazlı Can
Computer Engineering, Bilkent University _____

Date:10.01.2025

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Recep Fırat Çekinel

Signature :

ABSTRACT

MULTILINGUAL, MULTIMODAL AND EXPLAINABLE APPROACHES FOR AUTOMATED FACT-CHECKING PROBLEM

Çekinel, Recep Fırat
Ph.D., Department of Computer Engineering
Supervisor: Prof. Dr. Pınar Karagöz

January 2025, 116 pages

Automated fact-checking methods primarily rely on content-based approaches, utilizing deep neural networks to extract sophisticated features from text for prediction. However, the inherently black-box nature of these models makes their decision-making processes challenging to interpret. Another challenge for automated fact-checking models is their dependence on language-specific data, with limited multilingual datasets available for training. Moreover, the multimodal nature of fake posts—including text, images, and speech—presents an additional challenge. This thesis addresses automated fact-checking research, aiming to predict the veracity of claims while extending contributions to explainable solutions for fact-checking and sarcasm detection. We propose explainable models through multi-task learning and causal inference, evaluate cross-lingual transfer learning for low-resource languages, and examine how recent VLMs utilize text and image information for fact-checking. Our multi-task learning approach involves a T5-based encoder-decoder model trained for text summarization and veracity prediction, with generated summaries serving as explanations for predicted veracity labels. Moreover, a Turkish fact-checking dataset is released and experiments are conducted using transfer learning and machine trans-

lation to address data scarcity. In multimodality, we investigate VLMs' effectiveness in representing text and image information, finding that while multimodal embeddings generally enhance performance, discrete text-only and image-only models often outperform them. Lastly, we apply causal inference to text analysis, examining how sarcastic linguistic features and punctuation impact text popularity and leveraging clustering and topic modeling to uncover latent information on irony and popularity.

Keywords: fact-checking, explainability, cross-lingual learning, multimodality, causal inference

ÖZ

OTOMATİK DOĞRULUK KONTROLÜ PROBLEMİ İÇİN ÇOK DİLLİ, ÇOK MODLU VE AÇIKLANABİLİR YAKLAŞIMLAR

Çekinel, Recep Fırat
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Pınar Karagöz

Ocak 2025 , 116 sayfa

Otomatik doğruluk kontrolü yöntemleri, ağırlıklı olarak içerik tabanlı yaklaşımlara dayanmakta ve tahmin için metinden sofistike özellikler çıkarmak üzere derin sinir ağlarını kullanmaktadır. Ancak, bu modellerin doğası gereği kara kutu oluşu, karar verme süreçlerinin anlaşılmasını zorlaştırmaktadır. Otomatik doğruluk kontrolü modelleri için bir diğer zorluk, dil spesifik verilere olan bağımlılıklarıdır ve eğitim için mevcut çok dilli veri kümeleri sınırlıdır. Ayrıca, sahte gönderilerin metin, görseller ve konuşma gibi çok modlu yapısı ek bir zorluk teşkil etmektedir. Bu tez, otomatik doğruluk kontrolü araştırmalarını ele alarak iddiaların doğruluklarını tahmin etmeyi hedeflemekte ve doğruluk kontrolü ile alaycılık tespiti için açıklanabilir çözümler geliştirmeye katkı sağlamaktadır. Çok görevli öğrenme ve nedensel çıkarım yoluyla açıklanabilir modeller önermekte, düşük kaynaklı diller için çapraz dilli aktarım yoluyla öğrenimini değerlendirmekte ve son dönemdeki Görsel-Dil Modellerining (GDM) doğruluk kontrolü için metin ve görsel bilgiyi nasıl kullandığını incelemekteyiz. Çok görevli öğrenme yaklaşımımız, metin özetleme ve doğruluk tahmini için eğitilmiş T5 tabanlı bir kodlayıcı-çözücü modelini içermektedir; oluşturulan özetler, tahmin edilen

doğruluk etiketleri için açıklamalar olarak kullanılmaktadır. Ayrıca, Türkçe bir doğruluk kontrolü veri kümesi yayımlanmış ve veri yetersizliğini ele almak için transfer öğrenimi ve makine çevirisi kullanılarak deneyler gerçekleştirilmiştir. Çok modluluk bağlamında, GDM'nin metin ve görsel bilgiyi temsil etme etkinliğini araştırmakta ve çok modlu gömme yöntemlerinin genelde performansı artırmasına rağmen, yalnızca metin veya yalnızca görsel tabanlı modellerin sıklıkla daha iyi sonuç verdiğini bulmaktayız. Son olarak, metin analizinde nedensel çıkarımı uygulayarak alaycı dilsel özelliklerin ve noktalamanın metin popülerliği üzerindeki etkisini incelemekte ve ironi ve popülerlik hakkında örtük bilgileri ortaya çıkarmak için kümeleme ve konu modellemeyi kullanmaktayız.

Anahtar Kelimeler: doğruluk kontrolü, açıklanabilirlik, diller arası öğrenme, çoklu mod, nedensel çıkarım

To my family

ACKNOWLEDGMENTS

First, I would like to convey my deepest gratitude to my advisor, Prof. Pınar Karagöz. Her expertise and dedication have significantly shaped my dissertation and I am incredibly thankful for her guidance. I am also sincerely thankful to Dr. Çağrı Çöltekin for his invaluable insights which have greatly contributed to my research. Additionally, I appreciate Prof. Sinan Kalkan and Prof. Suat Özdemir for their advice and guidance.

I am immensely grateful to my dear friends, Kadir Cenk Alpay, Dr. Alperen Dalkıran, Güneş Sucu, Berk Kaya, Fahri Mert Ünsal, Mertalp Öcal, Dr. Nermin Samet and Dr. Samet Hiçsönmez, whose support and friendship have been a constant source of strength. Our numerous discussions have enriched my perspective and made this challenging journey more enjoyable.

Lastly, I want to extend my special thanks to my parents and my beloved sister Beyza Çekinel for their unconditional support and love. Their support made this achievement possible.

Finally, I would like to thank TUBİTAK for supporting my visit to the University of Tübingen through the 2214 International Doctoral Research Fellowship Program for 6 months and DAAD for their 7-month support through the Research Grant for Doctoral Students Program. Moreover, I also appreciate METU-ROMER and METU Image Lab for providing the computational resources. Parts of this research received the support of the EXA4MIND project, funded by the European Union´s Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS	
ATE	Average Treatment Effect
DAG	Directed Acyclic Graph
IFCN	International Fact-Checking Network
KNN	K-Nearest Neighbour
LDA	Latent Dirichlet Allocation
LR	Learning Rate
LLM	Large Language Model
MTL	Multi-task learning
NLP	Natural Language Processing
SCM	Structural Causal Model
SVM	Support Vector Machines
VLM	Vision Language Model

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Statement

There has been a significant increase in misinformation and disinformation on social media, with fake stories being fabricated to influence public opinion on various campaigns. Consequently, automated fact-checking studies have focused on assessing the truthfulness of claims using evidence documents.

These automated methods mostly depend on content-based approaches that utilize deep neural networks to extract sophisticated features from text for making predictions. However, deep neural networks are inherently black-box models, making their internal workings difficult to interpret. Providing explanations for their decision-making processes is vital, especially in critical decision-making contexts. This need has been emphasized by the recent EU AI Act legislation that mandates explanations for models used in decision-making. Additionally, new regulations require online platforms in the EU to ensure transparency in their reporting.

Another challenge for automated fact-checking models is their reliance on language-specific data. There are limited multilingual datasets available for training, making it essential to explore methods to learn from English training data and adapt to new domains and languages.

Social media platforms are increasingly becoming the primary source of news for many people. However, these platforms are susceptible to the rapid spread of fake stories, which can be used to manipulate public opinion [2]. Fabricated posts may include false text, images, videos, or speech content [3, 4, 5], designed to deceive

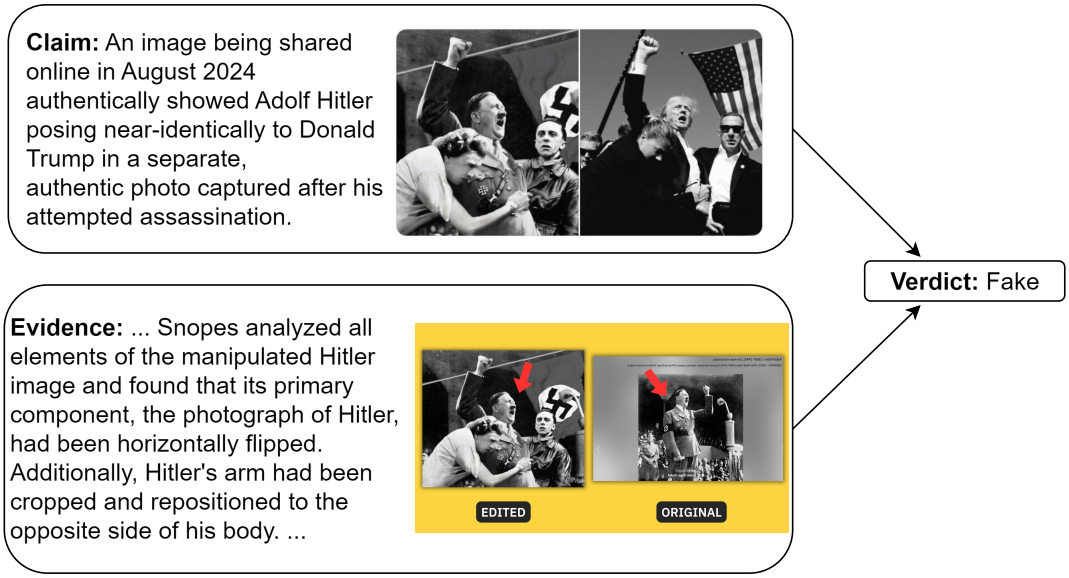


Figure 1.1: Example multimodal fact-checking from Snopes

social media users. Therefore, automated fact-checking systems should be able to consider information from different modalities [6]. For instance, on the Snopes website, a claim[1] about an edited image was proven to be fake by providing the original image and explaining how it was fabricated to manipulate public opinion about public figures. To verify the truthfulness of such content, it is essential to process both text and image information (see Figure 1.1).

This thesis details automated fact-checking research which involves predicting claims' veracity. Its contributions extend beyond fact-checking to offer broader explainable machine-learning solutions for fact-checking and sarcasm detection problems.

1.2 Proposed Methods and Models

Within the scope of this thesis, we tackle misinformation detection and sarcasm detection problems. First, we propose explainable models for these problems by implementing a multi-task model and leveraging causal inference. Additionally, we evaluate the viability of cross-lingual transfer learning for large language models that were massively pre-trained on English. Finally, we examine how recent VLMs utilize

[1] https://www.snopes.com/fact-check/hitler-trump-image-fake/

text and image information for fact-checking problem.

Multi-task learning: We propose a multi-task encoder-decoder model trained jointly for text summarization and veracity prediction. We formulate the explanation generation process for the model's veracity prediction as a text summarization problem. In other words, the generated summary serves as the explanation for the predicted veracity label. During training, the encoder is shared between both summarization and classification tasks. The decoder generates a summary, while the classification head takes the encoder representation as input and processes it through a feed-forward network for text classification. The overall loss is computed by taking the weighted sum of the summary and classification losses, using both static and dynamic loss weighting strategies.

Cross-lingual transfer learning: The state-of-the-art LLMs are massively pretrained in English and most annotated fact-checking datasets are in English as well. Therefore, we explore a cross-lingual transfer learning approach for low-resource languages, such as Turkish, in the context of fact-checking. Specifically, we aim to utilize an English fact-checking dataset to address the lack of data for Turkish. To achieve this, we collect a Turkish fact-checking dataset and conduct experiments with transfer learning by fine-tuning large language models and using machine translation. In addition to evaluating the feasibility of transfer learning, our results offer preliminary insights into the types of information, knowledge, or styles employed by automated fact-checking models.

Multimodality: VLMs utilize both text and image data to generate text responses. This thesis study investigates the application of VLMs for multimodal fact-checking, an area that has received limited attention. Specifically, it examines the effectiveness of VLMs in representing both text and image information. The proposed method extracts embeddings from the last hidden layer of selected VLMs and inputs them into a feed-forward network for multi-class veracity classification. The study first confirms the necessity of multimodal information using two multimodal fact-checking datasets. Then, it evaluates three recently released VLMs against separate text-only and image-only models.

Causal inference: Black-box models often rely on statistical correlations between features, which can introduce bias. To create robust systems, it is important to consider causality when estimating the data-generating process for a given task. In this thesis study, we investigate text-based causal inference for detecting irony and sarcasm. We also model latent confounders through unsupervised data analysis techniques, such as clustering and topic modeling. Our findings offer insights into causal explainability in irony detection.

1.3 Contributions and Novelties

Our contributions are as follows:

- We evaluate the effectiveness of multi-task training for veracity prediction and text summarization using a T5-based explainable model, finding that joint training enhances text classification performance in the Flan-T5 model with a slight decrease in summarization, while significantly improving summarization in the T5 model with less impact on classification. This work was also accepted to IEEE Big Data 2024 10^{th} Special Session on Intelligent Data Mining [7].

- We introduce a novel Turkish fact-checking dataset, comprising 3,238 claims with accompanying metadata and demonstrate that fine-tuning a large language model on this dataset significantly outperforms zero-shot and few-shot approaches, highlighting the value of language-specific for low-resource languages. This work was accepted to LREC-COLING 2024 [8].

- We propose a probing classifier pipeline using VLMs for multimodal fact-checking, finding that while multimodal embeddings generally improved performance over base VLM inference, discrete text-only and image-only models outperformed them. This work was accepted to COLING 2025 [9].

- We apply causal inference to text analysis to examine how sarcastic linguistic features and punctuation affect text popularity while also leveraging clustering and topic modeling to uncover latent information on irony and popularity. This work was accepted to DAWAK 2022 [10].

1.4 The Outline of the Thesis

The organization of this thesis is as follows:

- **Chapter 2:** Presents an extensive overview of the background information and existing literature review of the methodologies used in misinformation detection. The aim is to contextualize the current research landscape and highlight open problems that contribute to this thesis.
- **Chapter 3:** Presents a multi-task explainable misinformation detection model that combines veracity prediction and text summarization, where the summaries provide justifications for the model's predictions, utilizing a new architecture that integrates various neural models to enhance both fact-checking and explanatory tasks.
- **Chapter 4:** Includes releasing a Turkish fact-checking dataset, assessing the efficiency of transfer learning for low-resource languages like Turkish, demonstrating the superiority of fine-tuning a large language model with a Turkish dataset over few-shot learning and presents experimental results comparing zero- and few-shot prompt learning with fine-tuning on large language models, emphasizing the importance of using native data.
- **Chapter 5:** Examines the use of multimodal information (text and image) versus embeddings from text-only and image-only models for fact-checking and proposes a computationally efficient approach using the last hidden layer's representation from VLMs as input for a small feed-forward neural network to determine if recent VLMs can effectively leverage multimodal information for veracity classification.
- **Chapter 6:** Examines the causal effects of linguistic properties on irony and sarcasm detection, models latent confounders with K-Means clustering and LDA topic modeling and provides insights into causal explainability.
- **Chapter 7:** Summarizes the thesis findings, emphasizing the contributions to fact-checking through advanced machine learning techniques, and proposes potential future research directions.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

This chapter presents a comprehensive review of the fact-checking problem and existing literature relevant to the methodologies used in misinformation detection. The goal is to contextualize the current state of research and highlight the open problems that contribute to the thesis.

2.1 Misinformation Detection

Progresses in social networking and social media have not only made information more accessible but have also enabled the rapid spread of false information on these platforms [11]. As a result, disseminating fake stories has emerged as a powerful instrument for manipulating public opinion, as observed during the 2016 US Presidential Election [2]. Fake news can be described as media content that contains false information with the intent to mislead individuals [12, 13]. The goal of fake news detection is to evaluate the correctness of statements within the message content.

The traditional method of evaluating the correctness of a claim involves seeking the expertise of specialists who assess the claim by examining the available evidence. For instance, organizations like PolitiFact[1] and Snopes[2] rely on editors to validate the correctness of statements. However, this approach is both time-consuming and expensive. To address this issue, automated methods for fact-checking have emerged, intending to assess the truthfulness of claims while reducing the need for human intervention [14].

[1] https://www.politifact.com/
[2] https://www.snopes.com/fact-check/

Additionally, the explainability of fake news detection has gained attention. In general, biases inherent in AI systems, such as gender or race biases, can undermine model reliability [15]. Model explainability is about discovering the reasons behind the given model decision for a particular instance. While the black-box neural architectures dominate the fake news detection literature, various explainability approaches have been proposed to uncover AI-based models' reasoning.

Prior studies have revealed that fact-checking models often make predictions solely considering claim statements and ignore the evidence documents [16]. Additionally, such models make more mistakes on uncertainties such as "unproven" and "not enough information" cases. In situations where insufficient evidence exists to verify or refute a claim, it becomes crucial to determine the sufficiency of information for assessing correctness.

Like many other problems in NLP, the vast majority of available fact-checking resources released are primarily in English [17]. However, misinformation is not specific to content generated in English. Automated fact-checking systems are also needed for other languages, despite having much lower amount of expert annotated fact-checking data. Besides supervised data availability, the distribution of languages in pretraining data of state-of-the-art models also creates a big imbalance between English and other languages. Since creating large, manually annotated fact-checking data is a very expensive endeavor and finding the amount of unannotated data in languages other than English to (pre)train large language models are impractical (if not impossible), one promising solution is linguistic transfer: leveraging large datasets in English and cross-lingual transfer learning methods to build fact-checking systems for other, low-resource languages.

Fabricated social media posts may include false text, images, videos, or speech content [3, 4, 5], designed to deceive social media users. Therefore, automated fact-checking systems should be able to consider information from different modalities [6]. For instance, on the Snopes website, a claim [3] about an edited image was proven to be fake by providing the original image and explaining how it was fabricated to manipulate public opinion about public figures. To verify the truthfulness of such

[3] https://www.snopes.com/fact-check/zelenskyy-money-stacks/

content, it is essential to process both text and image information.

2.2 Causal Inference

Typical NLP models use statistical associations to make decisions and estimate the dataset's distribution using the training data. On the other hand, causal inference is an inverse problem that figures out the structural causal model of the data generating process, which leads to more robust and invariant models. Causal inference is about answering the counterfactual queries based on the intervention of interest. However, the counterfactual outcomes do not exist in the observational data in most cases. Therefore, the causal effect is the change of outcome variable Y by the intervention on treatment X when all other covariates are kept constant.

The initial step of causal inference represents the association between variables as Structural Causal Models (SCMs). The SCMs consist of directed acyclic graphs (DAGs) and a mathematical problem formulation. The variables are represented as nodes, and edges represent the causal relationship between variables.

Structural Causal Model: It consists of 3-tuples (U, V, E) where U denotes a set of exogenous variables (independent from the states), V denotes a set of endogenous variables (dependent to other states in the system) and they are connected by a set of structural equations, E, where each equation defines endogenous variables in terms of U and V.

After representing the causal model as a graph, interventions on a treatment can be expressed using Pearl's do-calculus notation [18]. Three rules of do-calculus which allow to simulate interventions on treatment to identify causal relationships in DAGs are summarized below:

- *Rule 1:* Insertion and deletion of observations

$$P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z), \text{ if W is irrelevant to Y}$$

- *Rule 2:* Action/observation exchange

$$P(Y \mid do(X), Z) = P(Y \mid X, Z), \text{ if Z blocks all back-door paths from X to Y}$$

- *Rule 3:* Insertion and deletion of actions

$$P(Y \mid do(X)) = P(Y), \text{ if there is no causal path from X to Y}$$

The first rule suggests that we can omit variables W if it is irrelevant to outcome Y. However, the second rule states that if variables Z blocks all backdoor paths from treatment X to Y, we must condition on Z. Finally, the third rule asserts that if there is no causal path from X to Y, we should not condition on X. A causal inference framework can estimate the counterfactual outcomes by making some assumptions that need to satisfy three criteria listed below:

- *Ignorability*: The treatment assignment and the counterfactual outcomes must be independent by randomizing the treatment assignment. However, for observational data, it is not feasible. Therefore, softer conditional ignorability criteria should be satisfied, which requires no unobserved confounders in the dataset.
- *Positivity*: For all covariates, the probability of receiving treatment must be greater than 0.
- *Consistency*: The outcome at unit i is only affected by the treatment at the same unit.

2.2.1 Causal Model Explainability

Texts are inherently high dimensional, and by encoding texts using language models, hidden factors such as topic, tone, and writing style can be discovered. BERT [19], a bi-directional transformer-based language model, had a breakthrough on NLP which outperformed previous models on many tasks with significant margins. However, Feder et al. [20] indicated that such models utilize statistical relationships while making decisions. Therefore, their predictions can be considered unreliable. Moreover, McCoy et al. [21] pointed out that these language models may fail when the data distribution of the test set changes significantly since these models rely on statistical associations. As a result, causal models are required to increase the models' generalization performance.

Secondly, the reasoning of any model can be evaluated with sensitivity and invariance tests. The former identifies how much minimal perturbation is necessary to switch the model's decision for the given sample. On the other hand, the latter determines whether a change in a causally unrelated feature impacts the model's decision. These tests can be valuable to interpret the model's robustness by feeding counterfactual inputs. Besides, Veitch et al. [22] stated that invariant models can perform better on different data distributions.

Language models such as BERT are not inherently explainable. According to Moraffah et al. [23] exploiting network artifacts such as attention weights is one approach to infer the decisions of a neural model. However, these approaches can only describe token-wise information. In addition, perturbating instances near decision boundary is another way of explainability [24, 25]. Yet, sentence-level estimates of such models may not be so successful [20]. In other words, these approaches may result in erroneous explanations to the decision-makers since they compute correlations between features [26, 27, 28].

In this context, causal models can generate counterfactual instances which can be used for interpretability [20]. For instance, a data sample's prediction can be compared with its counterfactual representative. More specifically, if a text contains a concept, its counterfactual will not include that concept, and their outputs can be compared to learn how the model makes decisions.

2.3 Related Work

2.3.1 Datasets

In recent years, numerous datasets have emerged for fact-checking and they can be categorized based on how claim statements are obtained. Some studies that create claim statements by extracting and manipulating content from source documents such as Wikipedia articles can be categorized as artificial claims [29, 30, 31, 32, 33]. These studies involve human annotators who systematically generate meaningful claims.

On the other hand, another approach involves collecting claims by crawling fact-

checking websites such as Politifact [34, 35] that primarily focuses on political claims and Snopes [36] that covers a broader range of topics. Additionally, some studies gather fact-checked claims from the Web [37, 38], specifically targeting domains like healthcare [39, 40], science [41], e-commerce [42]. Furthermore, Su et al. [43] introduced a hybrid dataset that includes both human-annotated and language model-generated claims.

Fact-checking datasets in languages other than English, and multilingual datasets are limited in comparison to English. FakeCovid [44] includes 5182 multilingual news articles related to COVID-19. DANFEVER [45], a Danish fact-checking dataset, comprises 6407 claims generated systematically following the FEVER [29] approach. Similarly, CsFEVER [46] features 3097 claims in Czech using a similar methodology. Additionally, CHEF [47] contains 10K claims in Chinese. Furthermore, CT-FCC-18 [48] contains political fact-checking claims in both English and Arabic, focusing on the 2016 US Election Campaign debates. X-Fact [49] comprises 31189 short statements from fact-checking websites across 25 languages. Lastly, Dravidian_Fake [50] consists of 26K news articles in four Dravidian languages.

The majority of existing datasets have concentrated on textual content for fact-checking. Nevertheless, some claims can benefit from the integration of various modalities, including images, videos and audio. Resende et al. [51] provide video, image, audio and text content from WhatsApp chats to detect the dissemination of misinformation in Portuguese. In addition both visual and textual information were utilized for fact-checking [52, 53, 54, 55, 56]. Additionally, MuMiN [57] incorporates the social context in the X platform (aka Twitter) and includes 12914 claims in 41 languages.

To the best of our knowledge, the only other fact-checking dataset that includes Turkish is X-Fact [49] which includes claims and evidence documents in 25 languages. Besides the differences in the size of the corpus, their Turkish data diverges from ours in a number of ways. Mainly, our focus in the corpus collection is richer monolingual data, rather than a large coverage of languages. The evidence documents in X-fact are through web searches, rather than crawling directly from the fact-checking site. Although there is some overlap in our sources, our data is also more varied in terms of fact-checking sites and topics of the claims. We also include short summaries pro-

vided in justifications and additional metadata. The summaries can be valuable for explainability in fact-checking [58, 39, 59, 60, 7]. In addition, a semi-automated method is applied to eliminate duplicate claims that we crawled from different sources.

2.3.2 Automated Fact-Checking

Automated fact-checking has been studied from data mining [12] and natural language processing [14, 17, 61] perspectives. The methods can be classified as content-based and context-based.

Zhou and Zafarani [13] further classify content-based methods as knowledge-based [62, 63] and style-based [64, 65, 66, 67]. Both approaches utilize news content to verify the veracity of a statement. While knowledge-based models assess statements by referencing their knowledge base, style-based methods typically prioritize assessing the lexical, syntactic and semantic attributes during verification.

Similarly, the authors categorized context-based methods as propagation-based [68, 69] and source-based [70, 71]. Both methods aim to capture social context to uncover the spread of information. While propagation-based models leverage interactions among users on social media by enhancing the interaction network with additional details like spreaders and publishers, source-based approaches rely on the credibility of sources which can also be employed to identify bot accounts on social media.

2.3.3 Explainable Fact-Checking

Kotonya and Toni [72] conducted a survey of the explainable fact-checking literature and classified the studies based on explanation generation approaches. These methods include exploiting neural network artifacts [73, 74, 75, 76, 77], rule-based approaches [78, 79, 80], summary generation [58, 39, 59, 60, 7], adversarial text generation [81, 82, 83], causal inference [84, 85, 86, 87], neurosymbolic reasoning [88, 89] and question-answering [90, 91].

Schmitt et al. [92] propose a framework for evaluating human-centric explanations for disinformation detection problem. The authors stated that free-text explanations

contribute to the non-expert individuals' performance. MADR [93] is a framework for generating faithful explanations utilizing LLM agents to refine explanations. JustiLM [94] employs a retrieval augmented generation (RAG) module for retrieving evidence and utilizes language models for explanation generation. And finally, the Climinator [95] framework parses claims into subclaims and uses specialized LLMs to evaluate these claims against credible sources. A mediator LLM synthesizes the verdicts of specialized LLMs.

The most related study in the literature was the E-BART model [60] that was trained for both classification and summarization by introducing a joint prediction head on top of the BART [96] language model. In other words, the encoder and decoder of the BART model are shared for both tasks. In contrast to this approach, this work incorporates the T5 Encoder as a shared module. For summarization, a T5 Decoder is trained while feed-forward layers are employed for classification. We also measured the effect of using two loss weighting strategies and evaluated the impact of instruction fine-tuning by switching the T5 model with the Flan-T5 [97] version.

Another related study in the literature was proposed by Atanasova et al. [58] who trained a joint veracity and summarization model (Explain-MT). The authors used DistilBERT [98] model's contextual embeddings for veracity prediction and text summarization and fed them to a cross-stitch layer [99]. This model generated extractive summaries which formed an outline using the evidence sentences. On the contrary, our multi-task model generate abstractive summaries which distill the information more conveniently. Additionally, according to Magooda et al. [100], multi-task learning could improve abstractive summarization performance in low-resource languages.

2.3.4 Cross-lingual Fact-Checking

Transfer learning approaches are limited for fact-checking. One approach in this field focuses on claim matching, aiming to link a claim in one language with its fact-checked counterpart in another language [101, 102]. Another approach focuses on out-of-domain generalization, involving the training of multilingual language models in a cross-lingual context [49, 103]. Besides, cross-lingual evidence retrievers can be employed to retrieve evidence documents in any language corresponding to a claim

made in a different language, thereby enhancing the cross-lingual fact-checking capabilities [104].

2.3.5 Text-Based Fact-Checking

Shared tasks such as FEVER [29], CLEF2018 [105] and AVeriTeC [106] evaluate fact-checking systems on textual claims. Although LLMs achieved high success rates on fact-checking with English data even in zero-shot settings [107], Zhang et al. [108] emphasize the need for language models that are specifically pre-trained on the target language. Similarly, Cekinel et al. [8] investigate cross-lingual transfer learning using LLMs. Additionally, FactLLaMA [109] incorporates external evidence during instruction-tuning to enhance the knowledge of LLMs. Moreover, MetaAdapt [110] focuses on cross-domain knowledge transfer with in-context learning. MiniCheck [111] verifies the factuality of synthetically generated claims against grounding documents. LLMs are also used for explanation generation [112, 94, 113] and neuro-symbolic program generation [88] for fact-checking. While these works primarily focus on enhancing models' knowledge, we aim to explore how they can leverage different modalities.

2.3.6 Multimodal Fact-Checking

While SpotFake+ [114] concatenates extracted text and image features for further processing through feed-forward layers, CARMN [115] fuses multimodal information using a cross-modal attention residual network. Pre-CoFactv2 [116] implements a multi-type fusion model that uses cross-modality and cross-type relations. COOLANT [117] implemented a contrastive learning based fusion method for image-text alignment. [118] incorporates the information extracted from the tweet graph with text and image embeddings for improving fake news detection. Liu et al. [119] examined the impact of audio in multimodal fact-checking by proposing a framework that fuses text, video and audio information with the cross-attention mechanism. Wang et al. [120] align news text with images by cross-modal attention model.

Geng et al. [121] propose an evaluation framework for VLMs that assesses the pre-

trained knowledge of these models in fact-checking without evidence. RAGAR [122] presents a RAG-based model that reframes the problem as question-answering for retrieved evidence pieces. MMIDR [123] trains a distilled model to generate explanations. SARD framework [124] applies multimodal semantic alignment to integrate multimodal network features. LVLM4FV [125] is an evidence-ranking approach and was evaluated on two benchmark datasets using LLMs and VLMs with zero-shot setting.

Although recent studies have focused on developing multimodal models for fact-checking using various fusion approaches, we aim to explore how effectively VLMs utilize different modalities. Geng et al. [121] also evaluated the robustness of recent VLMs for this problem by comparing the pre-trained knowledge of selected models and their prediction accuracy and confidence rates in zero-shot and few-shot settings. In contrast, we aim to leverage VLM representations by proposing a pipeline that trains a classifier using these embeddings. Furthermore, our primary focus is on utilizing multimodal information. In the experiments, we evaluate the intrinsic fusion of multimodal information against the extrinsic fusion of separate text-only and image-only representations.

2.3.7 NLP with Causality

Keith et al. [126] summarize the methods to adjust texts for causal inference. Moreover, Fong et al. [127] discuss the required assumptions to use latent features of text as treatment. In another study, they also use topic modeling to discover latent treatments in texts [128]. Moreover, Wood-Doughty et al. [129] address the challenges of using proxy treatments for causal inference.

Recently, Yang et al. [130] conduct a survey of existing causality extraction methods for texts. Moreover, Feder et al. [20] provide a review of the use-cases of text-based causal inference and discuss fairness, interpretability, and robustness aspects. Texts can be considered as treatment [1, 131], confounder [132, 126], outcome [133] and even mediator [134] settings. Sridhar et al. [135] examine the causal effect of tone on online debates. Koroleva et al. [136] propose a model to measure the similarity of pairs of clinical trial outcomes and reports semantically using BERT-based language

models.

There exist comprehensive studies that review models to explain black-box NLP models [137, 23, 20]. More recently, Chou et al. [28] also examine an in-depth review of the studies on model-agnostic counterfactual algorithms and argue that many such studies do not rely on causal theoretical formalism. Wang et al. [138] utilize a causal approach to exploit the attention weights of a sentiment classifier. Besides, perturbation-based approaches [24, 25] have been used for explanation. Another prominent and challenging text-based causal explanation method is counterfactual statement generation [139, 140, 141] which requires manipulating text in a meaningful manner. Therefore, instead of modifying the text itself, changing its representation has emerged by [142, 143]. Besides, Buyukbas et al. [144] and Cemek et al. [145] work on the same Turkish tweet dataset as in this chapter and examine the explainability of transformer architectures using two popular explainability tools, LIME [24] and SHAP [25] for irony detection task. Likewise, Hazarika et al. [146] propose the CASCADE model that utilizes both contextual and content information to improve the sarcasm detection performance significantly on SARC [147] dataset.

CHAPTER 3

EXPLAINING VERACITY PREDICTIONS WITH EVIDENCE SUMMARIZATION: A MULTI-TASK MODEL APPROACH

Multi-task learning (MTL) is a technique in machine learning to train similar tasks at the same time by leveraging their differences and commonalities [148, 149, 150]. Additionally, MTL allows data utilization as the model can transfer knowledge between tasks. Notably, the insights gained while learning one task can benefit other related tasks, leading to better generalization across tasks. Moreover, from the business point of view, deploying a single multi-task model may reduce the complexity of maintenance and resource requirements.

This chapter primarily focuses on designing a multi-task explainable misinformation detection model. To be more specific, a fact-checking model is trained on veracity prediction and text summarization tasks simultaneously. The generated summaries are derived from evidence documents and serve as justifications for the model's veracity prediction. Therefore, it should not be considered as a post-hoc explainability model. Figure 3.1 presents an example claim alongside our model's predictions. Based on supplementary information provided under the "Evidence" section, the claim has been verified by a human annotator. The gold summary was also written by human annotators while the abstractive summary was generated by a our multi-task model. The generated summary not only aligns with the veracity label but can be considered as an explanation of the model's reasoning behind its decision. More examples are provided in Appendix A.1.

The contribution of this work lies in the following:

- The use of multi-task learning for combining fact-checking and text summa-

Claim: Study says too many Americans still drink too much.
Evidence: "... The researchers found that 64 percent of men and 79 percent of women said they drank no alcohol at all that day, and another 18 percent of men and 10 percent of women drank within the recommended amounts. Nine percent of men said they had three to four drinks the day before and 8 percent of women said they drank two to three alcoholic beverages, the researchers said. The heaviest drinkers of all were the 8 percent of men who had five or more drinks, and 3 percent of women who had four or more. **"Overall the study confirms that rates of unhealthy alcohol use in the U.S. are significant**," said Jennifer Mertens, a research medical scientist at Kaiser Permanente Division of Research in Oakland, California, who was not part of the study. "
Gold Summary: On any given day in the United States, 18 percent of men and 11 percent of women drink more alcohol than federal guidelines recommend, according to a study that also found that 8 percent of men and 3 percent of women are full-fledged "heavy drinkers."
Our Summary: Americans are still drinking too much alcohol, even if they don't drink at all on any given day, according to a new study.
Veracity Label: TRUE

Figure 3.1: Sample instance from PUBHEALTH dataset with the gold justification and our model's summary.

rization tasks. The tasks, fact-checking and summarization, complement each other such that one does misinformation detection while the other explains the reason for the model's decision.

- Training a shared encoder and separate classification and summarization heads to perform both tasks simultaneously.

- Evaluating the performance of the proposed model on three benchmark datasets against the related studies.

The source codes are available at: https://github.com/firatcekinel/Multi-task-Fact-checking

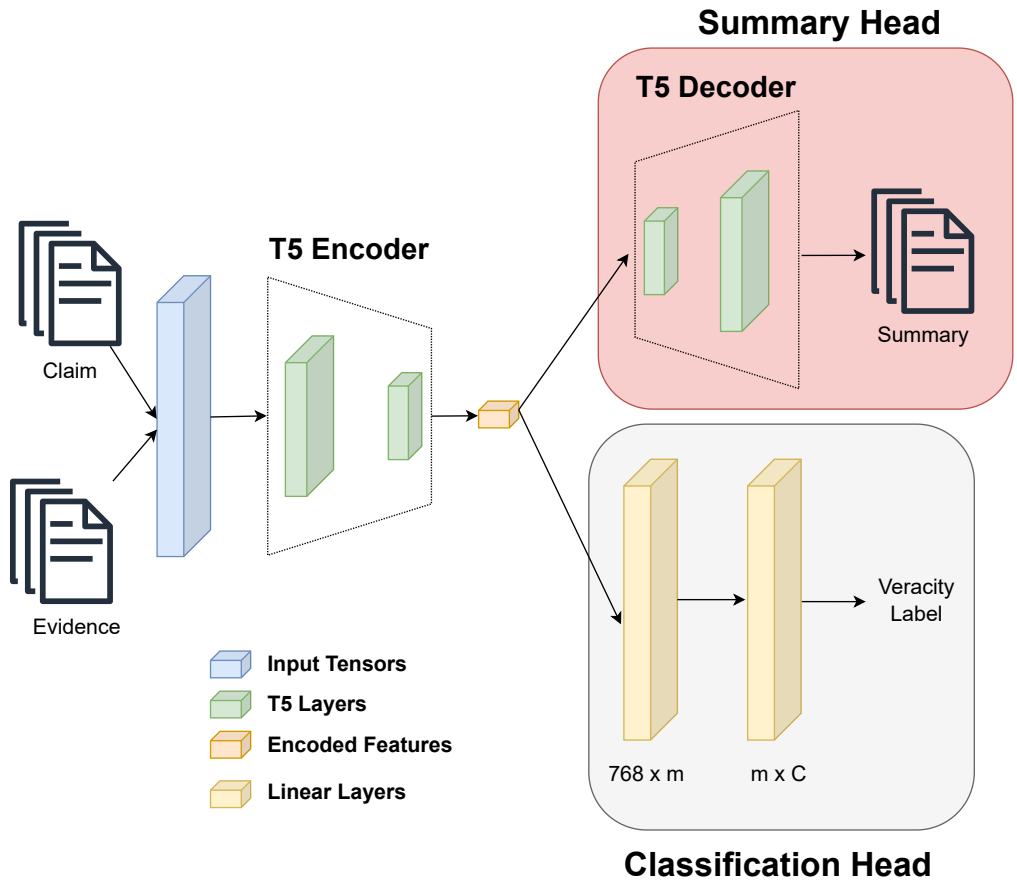


Figure 3.2: The multi-task model architecture

3.1 Method

In this chapter, a multi-task model that is based on the T5 [151] transformer is proposed. The model is trained on text summarization and veracity prediction tasks jointly. T5 transformer is an encoder-decoder model that converts each task to a text-to-text problem. Flan-T5 [97] model employs instructional fine-tuning to further improve the T5 model that is also utilized in the evaluation.

The model architecture is given in Figure 3.2. Both summarization and classification tasks share a T5 Encoder during training. At first, the T5 Encoder encodes the claim and evidence sentences in a latent space. Afterwards, the T5 Decoder produces a summary using the T5 Encoder's representation. Simultaneously, for the veracity prediction, the encoder's output is processed by two feed-forward layers respectively.

We employ the ReLU activation function and apply dropout between two linear layers and the sigmoid activation function after the second linear layer. Besides, the cross entropy loss is used for measuring summary and classification losses.

The overall loss is calculated by taking the weighted sums of the summary loss (w_s) and the classification loss (w_c) that is given in multi-task model loss Equation (3.1).

$$Loss = w_s \times Loss_{summ} + w_c \times Loss_{cl} \tag{3.1}$$

Two loss weighting strategies are employed: i) static loss coefficients and ii) uncertainty weighting. For the static loss coefficients, constant weights are set for the classification and summarization losses prior to training. To determine the optimal weights, grid search-based validation experiments are performed. In addition to the static loss coefficients, this paper also utilizes the uncertainty weighting strategy [152] that enables dynamic adjustment of the weights based on prediction confidence. Subsequently, the overall loss is calculated by taking the weighted sums of the summary loss and the classification loss.

3.2 Experimental Results

In this section, the proposed model was evaluated on three benchmark datasets. Note that we employed the T5-large and Flan-T5 large models in the Huggingface's transformer library and only the best results obtained during the validation experiments for each model are presented. Note that the experiments were conducted using Nvidia RTX A6000 GPUs.

3.2.1 PUBHEALTH Results

The PUBHEALTH [39] dataset consists of health-related claims with justifications written by journalists which were considered as gold explanations to evaluate the correctness of claims. Each claim was annotated as *True*, *False*, *Mixture* or *Unproven*. The training set consists of 9466 claims and 1183 claims exist in validation and test sets.

Table 3.1: Summarization results on PUBHEALTH

Model	Rouge-1	Rouge-2	Rouge-L
Oracle [39]	39.24	14.89	32.78
Lead-3 [39]	29.01	10.24	24.18
EXPLAINERFC-EXPERT [39]	32.30	13.46	26.99
T5 single-task	30.90	13.40	27.16
T5 multi-task	32.55	14.54	28.60
Flan-T5 single-task	**33.50**	**14.64**	**29.54**
Flan-T5 multi-task	32.38	14.03	28.41

In the experiments, the dropout and the learning rate were set to 0.1 and 1e-4 respectively with an Adam optimizer with a linear decay scheduler was employed. In addition, the hidden layer size (denoted as "m" in Figure 3.2) was determined as 128. The weights were assigned as follows: $w_{summary}$ and $w_{classification}$ were set to 0.5, $w_{mixture}$ to 2.5, and $w_{unproven}$ to 7.

To assess the summarization performance of the multi-task model, the ROUGE-N [153] and ROUGE-L [154] metrics were employed. These metrics were utilized to compare the proposed model against a baseline, an oracle, and summarization models implemented by Kotonya and Toni [39]. ROUGE-N measures the overlap of n-gram sequences between the ground truth and the given model's output. Likewise, the ROUGE-L metric captures the longest common co-occurring n-gram sequences.

Table 3.1 displays the summarization outcomes of the proposed models in comparison to the baseline and Oracle models. Lead-3 served as the baseline that utilized the first three sentences as a summary. Oracle was an extractive summary model that served as an upper bound. Additionally, EXPLAINERFC-EXPERT [39] was a single-task abstractive summary generator model that performed slightly better than T5 single-task model. On the other hand, the T5 multi-task model outperformed the state-of-the-art model in all Rouge metrics which implies that multitasking improved the summarization results for the T5 model. Note that the T5 single-task and the T5 multi-task models were almost identical to the model architecture given in Figure 3.2 but the classification head of the T5 single-task model was set to 0.

Table 3.2: Veracity results on PUBHEALTH

Model	Precision	Recall	F1-macro	Accuracy
BERT (rand. sentences) [39]	38.97	39.38	39.16	20.99
BERT (all sentences) [39]	56.50	56.50	56.50	56.40
BERT (top-k) [39]	77.39	54.77	63.93	66.02
SCIBERT [39]	75.69	66.20	**70.52**	69.73
T5 single-task	**78.24**	71.05	61.08	71.05
T5 multi-task	77.62	70.32	60.93	70.32
Flan-T5 single-task	74.80	73.56	61.39	73.56
Flan-T5 multi-task	76.46	**76.64**	65.18	**76.64**

Furthermore, the Flan-T5 multi-task model represents an instruction fine-tuned variant of T5 that performed slightly less effectively than the single-task Flan-T5 (for summarization), but both models outperformed the state-of-the-art model.

The results for veracity prediction using the precision, recall, F1-macro and accuracy metrics were presented in Table 3.2. The first two rows indicated the baselines. BERT (top-k) and SCIBERT models applied a sentence selection based on the sentences' semantic similarity with the claim sentences. For evidence selection, the authors employed the S-BERT [155] model. Therefore, we followed a similar approach and selected the top-5 evidence sentences and the claim statement as input for these models.

The results indicate that the Flan-T5 variant outperformed the T5-based models for veracity prediction but on the F1-macro metric the state-of-the-art SCIBERT model performed significantly better than the proposed models. The main reason for this difference can be attributed to the considerable imbalance in label distribution. For instance, the ratio of claims labeled as *Unproven* is approximately 3.2% while the *Mixture* cases constitute around 15.2% of the dataset.

The confusion matrix for the veracity prediction task 3.2 is given in Table 3.3 which revealed that the margins between the state-of-the-art model's and our models' F1-macro scores are attributed to the class distributions. More specifically, the dataset

Table 3.3: Confusion Matrix

Model		Unproven	False	Mixture	True	Accuracy
T5	Unproven	27	8	5	5	60.00
single	False	31	244	94	19	62.89
task	Mixture	17	41	131	12	65.17
	True	21	8	96	474	79.13
T5	Unproven	26	10	4	5	57.78
multi	False	31	236	106	15	62.43
task	Mixture	13	37	137	14	68.16
	True	15	17	99	468	78.13
Flan-T5	Unproven	25	14	1	5	55.56
multi	False	14	307	48	19	79.12
task	Mixture	9	61	87	44	43.28
	True	9	25	39	526	87.81

is highly imbalanced and despite boosting the *Unproven* and *Mixture* instances, the models suffered from the class imbalance problem. Moreover, another takeaway is that boosting the *Mixture* instances decreased the accuracy of *False* claims, particularly for T5 models.

3.2.2 FEVER Results

FEVER [29] is a benchmark dataset that includes 185K claims with related evidence documents from Wikipedia. The dataset was published for the FEVER shared tasks in 2018. For the fact-checking task, the claim statements were annotated as *Supports*, *Refutes* and *Not enough info*.

Since the FEVER test set did not contain the true labels, the multi-task model's veracity prediction performance was evaluated using the development set. To retrieve evidence documents, DOMLIN system [156] was employed. DOMLIN is a two-stage evidence retrieval system designed to enhance evidence recall. First, the document

retrieval module selects sentences that can be considered as potential evidence. Secondly, hyperlinks and the content of the hyperlinked pages are examined to uncover additional evidence. Additionally, the authors utilized BERT-base for evidence retrieval which was upgraded to ROBERTA-base [157] in the enhanced DOMLIN++ [59] version.

DOMLIN retrieved evidence documents for 17K out of the 20K claims in the development set, while labeling the remaining instances as "not enough info." With this supporting information, our multi-task model achieved an accuracy score of 76.18%. However, its Flan-T5-based counterpart outperformed it with a score of 80.44%. It's worth noting that the DOMLIN model [156] achieved an accuracy of 71.44%, DOMLIN++ [59] achieved 77.48%, and the E-BART [60] model reached an accuracy of 75.10% by utilizing the similar evidence retrieval method.

3.2.3 e-FEVER Results

The e-FEVER dataset [59] is a subset of the original FEVER dataset and consists of 67687 claims with evidence documents retrieved using the DOMLIN system. In addition to claims and evidence documents, the authors published the summaries using the GPT-3 model [158] for each claim. Hence, these summaries were leveraged as ground-truth explanations to compare our model's decision-making process with the GPT-3-based model.

The authors pointed out that the GPT-3-based model generated null summaries for certain claims. To address this issue, similar to Brand et al. [60], two variations of the dataset were utilized: *e-FEVER_Full* and *e-FEVER_Small*. The former contains all claims, while the latter excluded instances with null summaries. The *e-FEVER_Small* consists of 40702 instances. Moreover, the authors provided some examples labeled as *Not enough info* that could be either refuted or supported based on the provided evidence documents. Therefore, the binary veracity prediction performance of the multi-task model was measured by ignoring the *Not enough info* instances. Likewise, similar to Brand et al. [60] two variations of the multi-task model were trained: *T5-Small* and *T5-Full* where the former was trained on *e-FEVER_Small* and the latter was trained on *e-FEVER_Full*.

Table 3.4: Veracity and summarization results on e-FEVER

Model	Dataset	Acc. (w/o N.E.I)	Acc.	Rouge-1	Rouge-2	Rouge-L
E-BARTSmall [60]	eFever_Small	87.2	**78.2**	73.58	**64.37**	71.43
T5-Small	eFever_Small	**91.11**	74.75	74.00	63.64	72.78
T5-Small (uncertainty weighting)	eFever_Small	90.66	74.57	**74.46**	64.32	**73.19**
T5-Full (only summarization)	eFever_Full	-	-	65.94	57.53	65.09
Flan-T5-Full (only summarization)	eFever_Full	-	-	68.79	60.87	67.92
T5-Full (only classification)	eFever_Full	91.12	73.61	-	-	-
Flan-T5-Full (only classification)	eFever_Full	93.94	78.87	-	-	-
E-BARTFull [60]	eFever_Full	85.2	77.2	65.51	57.60	64.07
T5-Full	eFever_Full	90.91	75,26	68,16	59,96	67,26
T5-Full (uncertainty weighting)	eFever_Full	90.90	74,28	67,30	59,36	66,49
Flan-T5	eFever_Full	**94.36**	**79.91**	66.75	58.42	65.88
Flan-T5 (uncertainty weighting)	eFever_Full	93.94	79.02	**68.84**	**60.89**	**67.97**

After conducting several validation experiments, the best results were obtained on the e-FEVER dataset by setting the batch size to 4 and the hidden layer size (denoted as "m" in Figure 3.2) to 32. Moreover, we conducted experiments by employing both static loss weighting and uncertainty loss weighting strategies. For the static loss strategy, the weights were assigned as follows: $w_{summary}$ is set to 0.2 and $w_{classification}$ to 0.8.

Table 3.4 demonstrated the summarization and veracity prediction results on the e-FEVER dataset. To the best of our knowledge, only Brand et al. [60] reported results on this dataset. The first three rows indicated the models that utilized *e-FEVER_Small*

dataset. Both of the T5-based multi-task models performed slightly better than the E-BART Small model for summarization (except Rouge-2) and binary classification. However, E-BARTSmall achieved significantly higher accuracy (78.2%) than the proposed models in three-class classification.

Secondly, the baseline models were outlined starting from the fourth row to the seventh row which were trained specifically for either summarization or classification. Therefore, we did not report the classification results for the summarization model, and vice versa. Similarly, on *eFever_Full* the multi-task T5 models achieved higher binary classification accuracy and summarization scores but performed worse than the E-BARTFull model (77.2%) in multi-class classification. On the other hand, replacing T5 with the Flan-T5 version led to the highest accuracy scores in both binary and multi-class classification (94.36% and 79.91% respectively). Moreover, the Rouge scores of the T5 and Flan-T5 models were higher than the E-BART model on the *eFever_Full* dataset.

Furthermore, we also evaluated the impact of the loss strategy. To be more specific, we employed static loss weighting and uncertainty loss weighting which dynamically adapts the loss weights during training. According to the results, with uncertainty loss weighting the multi-task models performed slightly better on summarization but performed slightly worse on classification on both *eFever_Small* and *eFever_Full*. Overall, similar to the PUBHEALTH results, the multi-task models based on Flan-T5 demonstrated improved performance in classification through joint training. However, there was a slight decline in summary quality with multi-tasking. Conversely, T5-based models significantly improved on summarization with the aid of multi-tasking but decreased slightly in binary prediction accuracy.

CHAPTER 4

CROSS-LINGUAL LEARNING VS. LOW-RESOURCE FINE-TUNING: A CASE STUDY WITH FACT-CHECKING IN TURKISH

Cross-lingual learning has been studied in related problems such as hate speech detection [159], rumor detection [160], abusive language detection [161] and malicious activity detection on social media [162]. For fact-checking, Du et al. [163] proposed a model that jointly encodes COVID-19-related Chinese and English texts. Additionally, Raja et al. [50] employed joint training of English and Dravidian news articles and also applied zero-shot transfer learning by fine-tuning with English data and testing on Dravidian data.

Our primary aim in this chapter to test the viability of cross-lingual transfer learning approaches for fact-checking. We particularly focus on making use of data in English for fact-checking in Turkish for the cases of no or limited data availability. For this purpose, we collect a fact-checking data set for Turkish, and perform experiments with transfer learning through fine-tuning large language models and utilizing machine translation. Besides an assessment of the feasibility of transfer learning approaches, our results also provide some preliminary evidence for the type of information, knowledge or style, used in automated fact-checking models.

Our contributions can be summarized as:

- Releasing a Turkish fact-checking dataset obtained by crawling three Turkish fact-checking websites.[1]
- Assessing the efficiency of transfer learning for low-resource languages, with a specific emphasis on Turkish.

[1] https://github.com/firatcekinel/FCTR

- Presenting experimental results, comparing zero- and few-shot prompt learning and fine-tuning on large language models and underscoring the need to utilize a small amount of native data.

4.1 Data

Fact-checking datasets in both Turkish and English, are released by crawling Turkish fact-checking organizations and Snopes for English content. The significant similarity between the fact-checking domains of the Turkish websites and Snopes presents a valuable opportunity for transfer learning. In this chapter, various experiments are conducted to evaluate the necessity of collecting datasets in low-resource languages versus the effectiveness of transfer learning for these languages. Furthermore, we also conducted topic modeling to explore the latent topics within the datasets in Appendix B.1 and examined the potential content-based discrepancies between true and fake claims in Appendix B.2.

4.1.1 Dataset for Fact-Checking in Turkish (FCTR)

We crawled 6787 claims from the three Turkish fact-checking websites: Teyit[2], Dogrulukpayi[3] and Dogrula[4].

All are listed as fact-checking organizations on the Duke Reporters' Lab.[5] Dogrulukpayi and Teyit are also members of the International Fact-Checking Network (IFCN) which is a global community of fact-checkers. Our data collection process involved extracting *claim statements*, the corresponding *evidence* presented by the editorial teams, *summaries* providing justifications which are also written by the editors, *veracity labels*, *website URLs* and the *publication dates* of the URLs.

Claims retrieved from Teyit are summarized using the 'findings' section, which provides an overview of the evidence statements. Likewise, when it comes to claims sourced from Dogrula, the summary is derived from the final paragraph within the

[2] https://teyit.org/analiz
[3] https://www.dogrulukpayi.com
[4] https://www.dogrula.org/dogrulamalar
[5] https://reporterslab.org/fact-checking/

Figure 4.1: A fact-checked claim with multi-modal components [6]

'evidences' section, encapsulating the key findings. In the case of claims obtained from Dogrulukpayi, the dataset includes a dedicated paragraph following the rating section that encapsulates both the claim and the supporting evidence. This paragraph serves as the summary of these claims. Moreover, unique IDs were assigned to each claim in the dataset.

Claims were also marked as multi-modal if they contained keywords such as 'video', 'photo' and 'image' etc. This classification was made because we recognize that claims featuring such terms require verification not only of their textual content but also of any associated visual or video elements. For example, consider the fact-checked claim presented in Figure 4.1, which includes an image. In this claim, it was stated that the video shared on social media shows the moments when protesters in France set fire to the Alcazar Library in Marseille during the recent protests. The reviewer who gathered supporting information noted that 'According to inverse visual search results, the video is not from Marseille; it's from the Philippines. The building that caught fire is the Manila Central Post Office.' As a result, in order to verify such claims every aspect of evidences should be processed. Since our focus in this chapter is linguistic aspects of fact-checking, we do not make use of claims that require multimodal processing.

Last but not least, since the claims were collected from three distinct sources, we reviewed the claims to identify candidate duplicate claims. To accomplish this, the BERTScore metric [164] was employed that calculates a similarity score by analyzing the contextual embeddings of individual tokens within claim statements. We set the

[6] https://teyit.org/analiz/videodaki-yanginin- marsilyadaki-kutuphaneden-oldugu-iddiasi

Table 4.1: Veracity label counts in the FCTR dataset

Label	Sources	Counts
false	Dogrula, Teyit, Dogrulukpayi	2780
true	Dogrula, Teyit, Dogrulukpayi	203
mixed	Teyit	109
partially false	Dogrulukpayi	72
unproven	Teyit	37
half true	Dogrula	17
mostly false	Dogrula	14
mostly true	Dogrula	6

similarity threshold to 0.85 and execute the metric three times in data source pairs. Subsequently, a manual verification process was conducted to confirm whether the outputs from BERTScore indeed corresponded to duplicate claims.

After the preprocessing step, the dataset contains 3238 claims dating from July 23, 2016 to July 11, 2023. The value counts for each label are presented in Table 4.1. Furthermore, 742 claims of the final dataset were sourced from Dogrulukpayi, 525 claims were retrieved from Dogrula and 1971 fact-checked claims were gathered from Teyit.

4.1.2 Snopes Dataset

Snopes is an independent organization committed to fact-checking in English. They employ human reviewers who collect information about claims and write detailed explanations as justifications. It covers a broad range of topics, including politics, health, science, popular culture, etc. We collected claims along with their metadata including the justifications written by human annotators, veracity labels, website URLs and publication dates. We collected 6402 claims ranging from November 24, 1996 to August 17, 2023 and the label distribution is shown in Table 4.2. Even though Snopes covers a significantly wider date range than the FCTR, the majority of claims

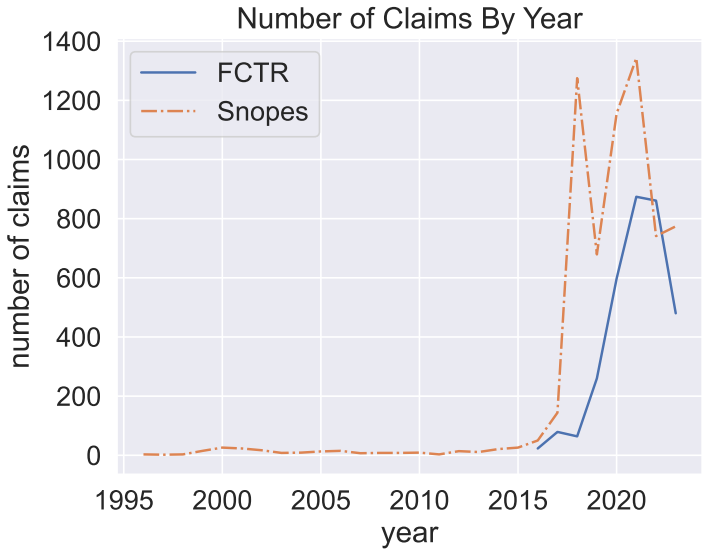


Figure 4.2: Number of claims by year in FCTR and Snopes datasets

are verified within the period from 2015 to 2023 as illustrated in Figure 4.2.

To the best of our knowledge, Snopes corpus was also crawled by Hanselowski et al. [36] and Augenstein et al. [37]. The reason why we re-collected the Snopes claims is that the previous corpus were released in 2019 but our FCTR corpus is up-to-date. Since we aim to evaluate the effectiveness of cross-lingual transfer learning and considering the potential overlap in fact-checking similar claims across both languages, we gathered the recent fact-checked claims in both English and Turkish.

4.2 Method

4.2.1 Model

In this chapter, we fine-tuned the LLaMA-2 [165] model for the veracity prediction task. Llama-2 is an open-source, auto-regressive transformer-based language model that was released by the Meta AI team. It has three variants, with parameter sizes of 7 billion, 13 billion, and 70 billion. Our main rationale for utilizing Llama-2 is

[7] 'other' encompasses the following labels: scam, outdated, misattributed, originated as satire, legend, research in progress, fake, recall, unfounded, legit

Table 4.2: Veracity label counts in the Snopes dataset[7]

Veracity Labels	Counts
false	2270
true	1467
mixture	588
miscaptioned	375
unproven	284
labeled satire	283
correct attribution	247
mostly false	237
mostly true	198
other	453

that it has a very large and almost up-to-date knowledge base. To be more specific, the pretraining data includes information up to September 2022, while the fine-tuning data is up to June 2023.

State-of-the-art language models comprise billions of parameters, demanding large GPU memory resources during fine-tuning for downstream tasks. Additionally, the deployment of such models in real-time applications has become increasingly impractical. Therefore, we adopted parameter-efficient fine-tuning and quantization to make the Llama-2 model fit within our GPU memory constraints without sacrificing information. First, LoRA [166] introduces a small number of additional parameters and updates their weights while keeping the original parameters frozen. Similarly, QLora [167] employs quantization to the frozen parameters to increase memory efficiency without a significant trade-off.

4.2.2 Instruction Prompting

Instruction tuning is a method that involves additional training of language models using template instruction-output pairs. It is shown that instruction tuning significantly improves the performance of large language models across a range of tasks [168].

This is because feeding such tuples to describe the task, allows it to better grasp the domain in question. Additionally, prompting was shown to be an effective way to describe models' reasoning steps by enabling the generation of coherent reasoning chains leading to the desired output [169].

Zero-shot prompting is a method of instructing a language model to generate predictions based on a provided prompt template, without the need for specific examples. During this decision-making process, language models can utilize both the knowledge that they acquired during pretraining and the template prompt. Zero-shot prompting proves particularly useful when you have fine-tuned a language model for a related task but lack labeled data for the specific task at hand. On the other hand, providing one or more examples from the intended task as prompts is referred to as few-shot prompting. By presenting these samples within the prompt, the model gains a better understanding of the desired output and its structure. Therefore, it often leads to superior performance compared to zero-shot prompting.

4.3 Experiments and Results

This section assesses the efficacy of transfer learning in the context of low-resource languages with a specific focus on Turkish. Note that only the best results achieved during the validation experiments for each model are presented.

4.3.1 Setup

The experiments were performed on two distinct datasets: *Snopes* and *FCTR*. Given the highly imbalanced nature of the Turkish fact-checking dataset, we conducted experiments on two variants of *FCTR*, namely *FCTR500* and *FCTR1000*. In the *FCTR500* dataset, all true claims along with 297 randomly sampled false claims were included. Conversely, in the *FCTR1000* dataset, 797 false claims were randomly sampled and combined with 203 true claims. *FCTR500* represents a balanced dataset, while *FCTR1000* serves as its imbalanced counterpart. Other labels were excluded because of their relatively low instance count and the varying labeling conventions within fact-checking communities for ambiguous cases such as partially true

```
### Instruction: Is the following statement "true"
or "false"?
### Input:
A series of photographs show the skeletal remains of
the biblical giant Goliath.
### Response:
false
```

Figure 4.3: Prompt template

and unproven claims. Similarly, when evaluating the language models on the Snopes dataset, we focused specifically on true and false instances. In both datasets, we randomly select 80% of the data for training, 10% for validation, and 10% for testing.

The SVM model [170] and the multilingual BERT (mBERT) model [171] were both trained on the same datasets with identical train-dev-test partitions as a baseline. For the SVM model, we used sparse word and n-gram features weighted by tf-idf. The training instances are weighted with inverse class frequency to counteract the class imbalance, particularly in the case of *FCTR100* trials. Similarly, we modified the cross-entropy loss function for the mBERT model. This adaptation took into account the inverse class ratios, causing the models to assign a higher penalty to the errors on the minority class compared to the majority class.

Prompt engineering played a critical role in the experiments. Various prompt formats were evaluated and the best results were achieved using the Alpaca prompt template [172], which is provided in Figure 4.3. The LLaMA-2 implementations in the Huggingface's transformers library were utilized language models in our transfer learning experiments. Although the LLaMA-2 language model was primarily pretrained on English data, we confirmed its proficiency in Turkish as well. Since it was pretrained on relatively recent data, we preferred LLaMA-2 in our experiments.

In the experiments, we used the SFTTrainer (from trl library) to fine-tune our models. While fine-tuning the LLMs cross entropy loss and Adam optimizer (paged_adamw_32bit)

with linear scheduler were employed. Additionally, we used a half-precision floating point format (fp16) to accelerate computations. Moreover, we applied parameter-efficient fine-tuning utilizing the QLoRA [167] method to fit the language models to Nvidia Quadro RTX 5000 and Nvidia RTX A6000 GPUs. The configuration included setting the dimension of the low-rank matrices (r) to 16, establishing the scaling factor for the weight matrices (lora_alpha) at 64, and specifying a dropout probability of 0.1 for the LoRA layers (lora_dropout).

4.3.2 Evaluation

In its prototypical use, fact-checking is very similar to many retrieval problems. We would like to identify a few non-factual texts (e.g., fake news) among (presumably) many factual documents (legitimate news). As a result, binary precision, recall and F1 scores considering non-factual texts as positive instances is a natural choice for evaluation. However, the datasets at hand provide an interesting challenge for evaluating fact-checking models. Since both classes are obtained from fact-checking organizations, most claims they care to consider are not factual.[8] Hence, the data sets at hand show a reverse class-imbalance compared to what we expect to observe in real use of such systems. As a result, for all experiments reported in this paper, we report F1-macro and F1-binary scores with respect to the 'false' class. The hyperparameter sweeps are performed to optimize the F1-macro score.

4.3.3 Results

Snopes Results: First of all, we conducted fine-tuning of the LLaMA and baseline models using the Snopes dataset. In all trials, input consisted solely of claim statements, without the inclusion of any supporting evidence. The results are summarized in Table 4.3. According to the results, the LLaMA-2 model with 70 billion parameters exhibited the best performance compared to other models. Since no supporting evidence was provided, the models were expected to rely on stylistic features for their

[8] Obtaining claims by other means may be a possible way to restore the class balance. However, such an approach also risks introducing spurious correlations with the veracity label (e.g., topic, style due to collection procedure).

Table 4.3: Veracity prediction on the Snopes data

Input	Model	F1-macro	F1-binary
claim 10-fold	SVM	0.651	0.709
claim	SVM	0.695	0.763
claim	mBERT	0.705	0.802
claim	LLaMA-7B	0.766	0.838
claim	LLaMA-13B	0.814	0.866
claim	LLaMA-70B	**0.826**	**0.890**

predictions. It is noteworthy that the SVM models learned purely from stylistic features. Nevertheless, a substantial performance gap exists between the SVM and the LLaMA-2 models. This margin could be attributed to the pretrained knowledge embedded in LLaMA-2 models. Moreover, the larger LLaMA-2 models outperformed LLaMA-7B, suggesting that LLaMA-13B and LLaMA-70B leverage their knowledge better than their smaller variant.

FCTR Results: Table 4.4 and Table 4.5 present the fine-tuning results on the *FCTR500* and *FCTR1000* datasets respectively. According to the findings, when using only the claim statement as input, the SVM model which bases its predictions solely on stylistic features achieved the highest F1-macro score on the *FCTR500* and *FCTR1000* datasets. While evaluating with claim statements only, on *FCTR1000* dataset, we fine-tuned the LLaMA models on the Snopes dataset for two epochs initially and continued fine-tuning on the *FCTR1000* dataset for one epoch to achieve the best results. Besides, the class weights of the cross entropy loss function of the multilingual BERT model were adjusted according to the class proportions inversely to get the best result.

Furthermore, when both the claim statement and the summary (which summarizes the evidence provided by crowd workers) were given as input, the LLaMA-13B model reached a superior 0.89 and 0.828 F1-macro scores on *FCTR500* and *FCTR1000* datasets respectively and 0.923 and 0.947 F1-binary scores respectively. These scores were substantially higher compared to training the model with claims alone. The rea-

Table 4.4: Fine tuning on the FCTR500 data

Input	Model	F1-macro	F1-binary
claim 10-fold	SVM	0.682	0.610
claim	SVM	**0.714**	0.709
claim	mBERT	0.653	0.750
claim	LLaMA-7B	0.632	0.765
claim	LLaMA-13B	0.635	0.679
claim	LLaMA-70B	0.649	**0.783**
+summary	mBERT	0.752	0.861
+summary	LLaMA-13B	**0.890**	**0.923**

Table 4.5: Fine tuning on the FCTR1000 data

Input	Model	F1-macro	F1-binary
claim	SVM	**0.671**	0.842
claim	mBERT	0.518	0.797
claim	LLaMA-7B	0.561	**0.864**
claim	LLaMA-13B	0.642	0.839
+summary	mBERT	0.729	0.902
+summary	LLaMA-13B	**0.828**	**0.947**

son why we incorporated summaries as input was to examine whether this additional information improves the models' capabilities. Notably, the LLaMA models have limited proficiency in Turkish and we observed poor performance when solely presented with claim statements.

4.3.4 Assessing the Impact of Number of Training Instances

In this experiment, we examined the influence of varying training data quantities on model performance. We maintained consistency by utilizing the identical test set

Table 4.6: Impact of number of inputs on the FCTR500 data

Model	Input	F1-macro	F1-binary
LLaMA-7B	50 claims	0.566	0.644
LLaMA-7B	100 claims	0.570	0.716
LLaMA-7B	200 claims	0.576	0.677
LLaMA-7B	300 claims	**0.649**	**0.783**
LLaMA-7B	400 claims	0.632	0.765

employed in the previous experiment given in Table 4.4. Table 4.6 illustrates the consequences of manipulating the quantity of training data when employing the LLaMA-7B model. According to the results, as the number of training instances increases, the F1-macro score exhibits gradual improvement. However, when we employed 300 and 400 training instances, the model's performance remained almost constant, with both cases yielding remarkably similar results with only a single instance having a label change in the negative direction. This observation suggests that beyond a certain threshold, additional training instances may not provide substantial performance gains, highlighting the presence of a saturation point in the learning curve.

4.3.5 Cross-Lingual Transfer Learning

Zero-shot learning and few-shot learning can be achieved by providing prompts to large language models. In the zero-shot setting, no specific instances are provided for the given task. Instead, the model makes predictions based solely on the provided instructional prompts and input statements. In contrast, in the K-shot setting, K instances for each class along with their labels are included in the input prompt. This approach enables the model to gain a better understanding of the task's intention and the desired answer format. We evaluated the effectiveness of transfer learning on two distinct datasets: *FCTR500*, which is more balanced, and *FCTR1000*, which is imbalanced. Note that in the experiments, we employed the models that were fine-tuned on the Snopes dataset with the corresponding results provided in Table 4.3.

Moreover, we conducted transfer learning experiments by repeating few-shot settings five times and reported the average scores along with the standard errors. According to Table 4.7 and Table 4.8, few-shot learning appears to be beneficial for the LLaMA variants. In other words, providing sample instances within prompts slightly enhanced their performance. However, fine-tuning LLaMA language models with Turkish data resulted in a substantial improvement in the F1-macro score. For instance, on the *FCTR1000* dataset, while few-shot learning achieved the highest F1-macro score of 0.560 (in Table 4.8), fine-tuning with Turkish data boosted all LLaMA variants to F1-macro score of 0.642 (in Table 4.5).

4.3.6 Neural Machine Translation

Neural machine translation is an approach that employs deep learning models to translate a text from a source language to a target language [173]. The transformer-based generative large language models are pretrained massively in English. Therefore, their performance in other languages may not be equally impressive. To tackle this challenge, we conducted translations of the Turkish fact-checking dataset into English utilizing the ChatGPT API. Table 4.9 presents the veracity detection results on the translated data. Note that we employed the models fine-tuned on the Snopes dataset.

The results suggest that employing translated claims led to higher success rates for LLaMA models compared to the few-shot prompting approach. However, the success rate of mBERT was not positively influenced by translation. This phenomenon may be attributed to the differences in pretraining data between LLaMA models and mBERT. To be more specific, the LLaMA models were massively trained on English corpora, while the pretrained data for mBERT might exhibit a more uniform language distribution.

Additionally, we annotated the test set of *FCTR500* data based on claim statements, marking them as either "local" or "global". Claims that specifically related to Turkiye were marked as "local" claims, while claims with broader implications were labeled as "global". This categorization was done to assess the impact of the LLaMA model's pretrained knowledge on the claim category. We expected that the model would per-

Table 4.7: Transfer learning on the FCTR500 data

Input	Model	F1-macro	F1-binary
zero shot	mBERT	0.550	0.667
zero shot	LLaMA-7B	0.488 ∓ 0.026	0.577 ∓ 0.027
1-shot	LLaMA-7B	0.536 ∓ 0.006	0.742 ∓ 0.009
2-shot	LLaMA-7B	0.545 ∓ 0.035	0.632 ∓ 0.045
3-shot	LLaMA-7B	0.577 ∓ 0.011	0.642 ∓ 0.029
4-shot	LLaMA-7B	0.538 ∓ 0.021	0.609 ∓ 0.024
5-shot	LLaMA-7B	0.533 ∓ 0.021	0.647 ∓ 0.022
zero shot	LLaMA-13B	0.498 ∓ 0.014	0.699 ∓ 0.006
1-shot	LLaMA-13B	0.489 ∓ 0.026	0.683 ∓ 0.023
2-shot	LLaMA-13B	0.530 ∓ 0.028	0.689 ∓ 0.019
3-shot	LLaMA-13B	0.482 ∓ 0.022	0.670 ∓ 0.028
4-shot	LLaMA-13B	0.529 ∓ 0.036	0.638 ∓ 0.028
5-shot	LLaMA-13B	0.514 ∓ 0.013	0.632 ∓ 0.007
zero shot	LLaMA-70B	0.527 ∓ 0.042	$\mathbf{0.773} \mp 0.016$
1-shot	LLaMA-70B	0.507 ∓ 0.036	0.766 ∓ 0.018
2-shot	LLaMA-70B	0.539 ∓ 0.021	0.754 ∓ 0.013
3-shot	LLaMA-70B	0.492 ∓ 0.030	0.692 ∓ 0.023
4-shot	LLaMA-70B	0.542 ∓ 0.021	0.709 ∓ 0.014
5-shot	LLaMA-70B	$\mathbf{0.585} \mp 0.017$	0.709 ∓ 0.023

form better on global claims, given the possibility that it might have pretrained information related to such claims from the web. The results indicate that using the LLaMA-13B model, the average F1-macro for local claims was 0.520 ∓ 0.036 while the average F1-macro score for global claims was 0.582 ∓ 0.056. However, using the LLaMA-7B model, we obtained the average F1-macro scores of 0.567 ∓ 0.017 for local claims and 0.541 ∓ 0.015 for global claims. The results imply that the higher F1-macro score for global claims with the larger LLaMA model may be attributed to its pretraining knowledge that should be addressed in further research.

Table 4.8: Transfer learning on the FCTR1000 data

Input	Model	F1-macro	F1-binary
zero shot	mBERT	0.529	0.736
zero shot	LLaMA-7B	0.479 ∓ 0.019	0.647 ∓ 0.018
1-shot	LLaMA-7B	0.501 ∓ 0.017	0.857 ∓ 0.013
2-shot	LLaMA-7B	0.518 ∓ 0.010	0.706 ∓ 0.006
3-shot	LLaMA-7B	0.501 ∓ 0.010	0.691 ∓ 0.024
4-shot	LLaMA-7B	0.512 ∓ 0.023	0.694 ∓ 0.024
5-shot	LLaMA-7B	0.502 ∓ 0.030	0.690 ∓ 0.048
zero shot	LLaMA-13B	0.502 ∓ 0.011	0.803 ∓ 0.006
1-shot	LLaMA-13B	0.550 ∓ 0.016	0.811 ∓ 0.014
2-shot	LLaMA-13B	0.539 ∓ 0.033	0.788 ∓ 0.020
3-shot	LLaMA-13B	0.533 ∓ 0.017	0.763 ∓ 0.016
4-shot	LLaMA-13B	0.537 ∓ 0.010	0.758 ∓ 0.010
5-shot	LLaMA-13B	0.533 ∓ 0.029	0.737 ∓ 0.021
zero shot	LLaMA-70B	0.521 ∓ 0.018	$\mathbf{0.865} \mp 0.002$
1-shot	LLaMA-70B	0.528 ∓ 0.011	0.858 ∓ 0.011
2-shot	LLaMA-70B	$\mathbf{0.560} \mp 0.033$	0.841 ∓ 0.012
3-shot	LLaMA-70B	0.536 ∓ 0.023	0.806 ∓ 0.018
4-shot	LLaMA-70B	0.520 ∓ 0.019	0.808 ∓ 0.016
5-shot	LLaMA-70B	0.521 ∓ 0.018	0.778 ∓ 0.015

Furthermore, we employed Opus-MT's [174] *opus-mt-tc-big-en-tr* model to translate the Snopes dataset into Turkish and subsequently fine-tuned the language models using the translated Snopes' claims. This experiment was conducted to examine the impact of translating an English dataset into a low-resource language, specifically Turkish, on model performance. The fine-tuned models were then evaluated on the test splits of *FCTR500* and *FCTR100* to maintain consistency with the other experiments. According to Table 4.10, the F1-macro scores slightly decreased compared to the results presented in Table 4.9 when translating to a low-resource language.

Table 4.9: Turkish to English machine translation results

Dataset	Model	F1-macro	F1-binary
fctr500	mBERT	0.561	**0.789**
fctr500	LLaMA-7B	**0.576** $\mp$ 0.014	0.782 $\mp$ 0.007
fctr500	LLaMA-13B	0.567 $\mp$ 0.018	0.739 $\mp$ 0.013
fctr500	LLaMA-70B	0.571 $\mp$ 0.015	0.771 $\mp$ 0.007
fctr1000	mBERT	0.485	0.840
fctr1000	LLaMA-7B	0.524 $\mp$ 0.011	0.847 $\mp$ 0.003
fctr1000	LLaMA-13B	0.573 $\mp$ 0.013	0.879 $\mp$ 0.004
fctr1000	LLaMA-70B	**0.581** $\mp$ 0.012	**0.883** $\mp$ 0.003

Table 4.10: English to Turkish machine translation results

Dataset	Model	F1-macro	F1-binary
fctr500	mBERT	0.532	**0.757**
fctr500	LLaMA-7B	0.523 $\mp$ 0.019	0.630 $\mp$ 0.023
fctr500	LLaMA-13B	0.544 $\mp$ 0.018	0.708 $\mp$ 0.006
fctr500	LLaMA-70B	**0.553** $\mp$ 0.025	0.725 $\mp$ 0.022
fctr1000	mBERT	0.474	0.826
fctr1000	LLaMA-7B	0.481 $\mp$ 0.023	0.705 $\mp$ 0.020
fctr1000	LLaMA-13B	0.552 $\mp$ 0.044	0.800 $\mp$ 0.024
fctr1000	LLaMA-70B	**0.556** $\mp$ 0.018	**0.832** $\mp$ 0.011

Fine-tuning on translated data involves certain considerations. To be more specific, despite the state-of-the-art machine translation models accurately translating content, it might not be always feasible to maintain all context after translation. Additionally, since the current language models have a better understanding of English, it is an expected outcome that they would exhibit better performance on data translated from Turkish to English. Likewise, the results suggested that collecting native data for low-resource languages (Turkish for this case) is still required to ensure the development

of successful models.

4.4 Discussion

The main objective of this chapter is to test the possibility and the extent of making use of a large amount of fact-checking data and large language models that were heavily pretrained in English for fact-checking in other languages with much less labeled data, and much smaller pretraining data for large language models. We focus on Turkish as a low-resource language for this task. Although focusing on a single familiar language allows us to curate a better fact-checking corpus, and perform more meaningful error analysis, our approach is applicable to many languages. Results are likely to differ based on typological similarity of the languages in question, as well other factors like geographical proximity and cultural similarity of the communities that speak the language.

Our experiments demonstrate some small gains from the high-resource language in zero-shot and few-shot settings, where few-shot learning shows slight improvement over zero-shot. The results in Table 4.7 and Table 4.8 shows a small but consistent increase in F1-macro scores when a few examples are included. The benefit of more few-shot examples is unclear, however. The same is true for making use of machine translation from low-resource language to high-resource language. The test instances translated to English labeled by the models trained on English data clearly better than an uninformed system. Even a small amount of training data provides better results than zero- or few-shot approaches.

Another interesting outcome of our results is the success of small models that rely only on surface cues on the FCTR data. There are no obvious latent variables (e.g., authors, source websites) that can identify the veracity label of short claim texts. This means some relevant information is available on the surface features. However, the large language models surpass the simple ones on English with a large margin (see Table 4.3). This may indicate both the help of the linguistic and perhaps factual information brought by these models.[9] However, most probably the comparatively

[9] A potential problem here is these models may have the full fact-checking report for the test instances, including the clearly stated verdict in their pretraining data.

smaller Turkish data during pretraining is possibly a factor in low scores of LLaMA with fine-tuning with Turkish (Tables 4.4 and 4.5).

In the majority of the experiments, only the claim statements were employed as input, since this is a more realistic scenario as individuals typically seek to assess the truthfulness of a claim before spending time gathering additional information. We also include evidence statements as input in some experiments, which show a clear benefit in providing additional information. However, evidence retrieval is also a challenging problem in fact-checking (which falls beyond the scope of this chapter). A further problem with providing evidence may be discouraging the model from leveraging its pretrained knowledge while making decisions.

CHAPTER 5

MULTIMODAL FACT-CHECKING WITH VISION LANGUAGE MODELS: A PROBING CLASSIFIER BASED SOLUTION WITH EMBEDDING STRATEGIES

A vision language model (VLM) consists of an image encoder, a text encoder and a mechanism such as contrastive learning [175] and cross attention [176] to fuse text and image information. By this way, the model leverages the text and visual information while generating a response text. VLMs consist of billions of parameters and fine-tuning these models requires significant computational resources. Although parameter-efficient fine-tuning approaches [177, 178] have proven to be very effective for large language models, VLMs do not scale well horizontally. Consequently, such VLMs cannot be fine-tuned with moderate batch size and sequence length on a single GPU for problems like fact-checking that requires long text inputs.

Instead of fine-tuning, probing classifiers are trained on the representations of a pre-trained model [179] to predict linguistic features such as dependency parsing [180] and POS tagging [181]. A key advantage of probing classifiers is their ability to assess how well the pre-trained model has captured linguistic properties. In this chapter, we aim to evaluate how VLMs leverage both text and images for the fact-checking task by training a probing classifier. The following research questions are addressed in the paper.

RQ1: Validating the need for multimodality: Does incorporating multimodal data improve performance in the fact-checking task or are text-only models sufficient?

RQ2: Leveraging multimodal content: How effectively do VLMs utilize both text and image information to enhance fact-checking performance?

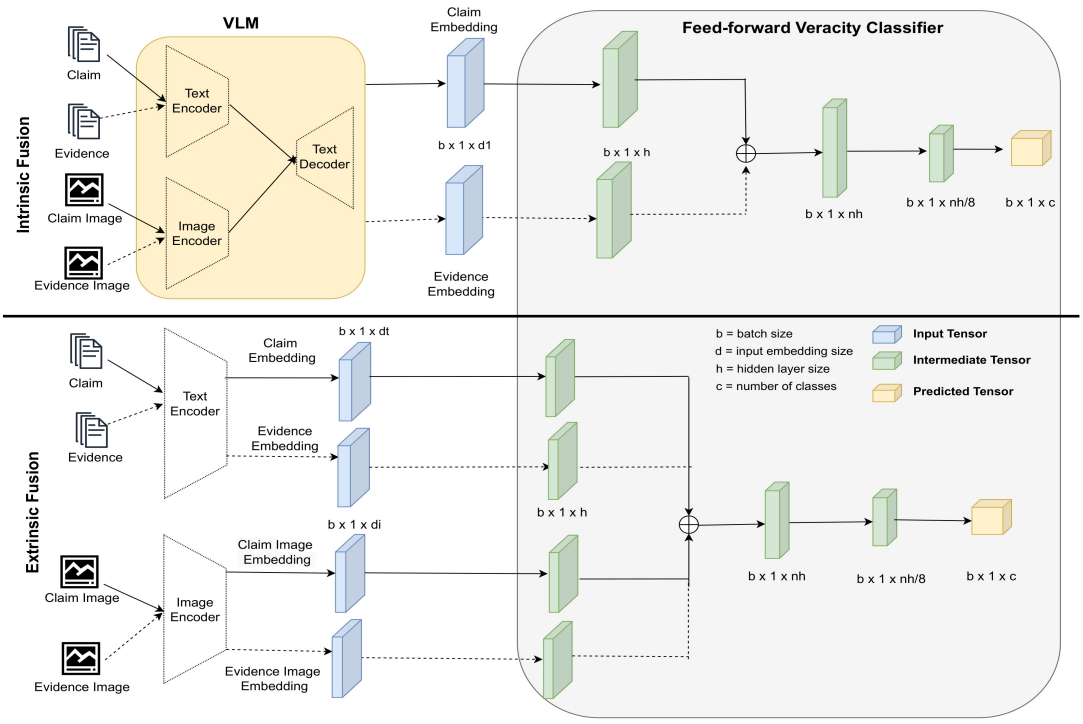


Figure 5.1: Overview of our probing fact-verification classifier. ReLU activation is applied after each linear layer with dropout for better generalization. The dashed lines indicate optional embeddings. In other words, evidence text and evidence image representations are optional in this pipeline.

RQ3: Evaluating probing classifiers: How does a probing neural classifier compare to baseline models in the context of the fact-checking task?

This chapter proposes a probing classifier that involves extracting the last hidden layer's representation and using it as input for a neural network. By introducing this pipeline, we aim to elaborate on the utilization of multimodal information, text and image, compared to embeddings extracted from discrete text-only and image-only models for the fact-checking problem. The source code is available at the following GitHub repository[1].

[1] https://github.com/firatcekinel/Multimodal-Fact-Checking-with-Vision-Language-Models

5.1 The Proposed Method

5.1.1 Feed-Forward Veracity Classifier

We introduce a probing classifier to examine the efficiency of multimodal embeddings compared to separate embeddings extracted from text-only and image-only models for veracity prediction. The VLM embeddings fuse text and image modalities intrinsically but distinct text and image encoder embeddings are fused extrinsically by the probing classifier as illustrated in Figure 5.1.

First, the last hidden layer representation is extracted from a VLM or a text/image encoder. The neural classifier either receives the VLM representation or embeddings from the corresponding text encoder and image encoder, then predicts veracity classes. If multiple input tensors are fed to the neural classifier, they are processed by a linear layer and after the first layer, all tensors are resized to a "hidden_size" — a hyper-parameter determined by validation experiments — and then concatenated. We concatenate after the first layer because the text and image embedding sizes vary significantly. To utilize both types of information equally, we resize these embeddings to the same dimension and concatenate them afterward. On the other hand, if only the VLM embedding is given to the network as input, two linear layers process the tensor sequentially without any concatenation.

In both of the probing classifier architectures, we implement a weighted cross-entropy loss, with weights determined by inverse class ratios to penalize the majority class more. Since PyTorch's cross-entropy loss implementation combines softmax with negative log-likelihood loss, the output tensor predicts class probabilities. Consequently, the classifier predicts the class with the highest probability for a given instance.

5.1.2 Models

The primary goal of this chapter is to examine whether merging image and text information provides gains for the fact-checking problem. To this end, we selected three multimodal models with different fusion mechanisms, as explained below.

Qwen-VL [182] is a multimodal model introduced by Alibaba Cloud. Qwen-VL is based on the Qwen-7B [183] language model and Openclip's ViT-bigG [184] vision transformer. The model leverages both modalities through a cross-attention mechanism. Information from the vision encoder is fused into the language model using a single-layer cross-attention adapter with query embeddings optimized during the training phase. In this chapter, we employed *Qwen-VL-Chat-Int4* checkpoint which was the 4-bit quantized version.

Idefics2 [185] is a general-purpose multimodal VLM introduced by Huggingface. It is based on the Mistral-7B [186] language model and SigLIP's vision encoder [187] (SigLIP-So400m/14). The model employs a vision-language connector that takes the vision encoder's representation as input, using perceiver pooling and MLP modality projection. After these operations, the image information is concatenated with the encoded text representation and fed into the language model decoder.

PaliGemma [188] is introduced by Google and is based on the Gemma-2B [189] language model and SigLIP's vision encoder [187] (SigLIP-So400m/14). Since Gemma-2B is a decoder-only language model, the vision encoder's representation is fed into a linear projection, concatenated with text inputs, and then fed into the Gemma-2B language model for text generation. In this chapter, we employed *paligemma-3b-mix-448* checkpoint that was fine-tuned on a mixture of downstream tasks.

5.1.3 Datasets

Mocheg [55] consists of 15K fact-checked claims from Politifact and Snopes. These websites employ journalists to verify claims who collect evidence documents and write ruling comments. The Mocheg dataset includes both text and image evidence which were crawled from the reference articles linked on the fact-checked claims' webpages. In cases where multiple evidence images were available for a claim, some collected images were found to be irrelevant. Therefore, for the experiments, only the first image was used as the evidence image.

Factify2 [190] is a challenge dataset containing 50K claims. The authors collected true claims from tweets by Indian and US news agencies and false claims from fact-

checking websites. They scraped text and image evidence from external articles and also collected claim images from the headlines of the claims. The fact-verification task was reformulated as an entailment problem where claims were annotated to indicate whether the claim text and image were entailed by the evidence text and image.

5.2 Experiments

We conducted experiments on compute nodes with 4x40GB Nvidia A100 GPUs. While evaluating the models on the datasets, we ignore the instances that have missing text evidence or images. For the Mocheg dataset, we used the original train-dev-test splits. The dataset has three labels *"supported"*, *"refuted"* and *"not enough info (NEI)"* and we used the labels as it is.

Regarding the Factify2 dataset, since the labels in the test set were unavailable, the original validation data was kept for testing. Instead, we randomly selected 10% of the training set for validation but kept the same percentages of classes in each split. Similar to Tahmasebi et al. [125], we reduced the original five labels to three classes: *Support* (Support_Multimodal & Support_Text), *Refute* and *Not enough info* (Insufficient_Multimodal & Insufficient_Text) to evaluate the proposed approach.

During the training of the probing classifier using the embeddings, validation experiments were conducted through grid search within the parameter space detailed below. Note that only the best parameter settings are presented in Appendix C.1. Last but not least, we reported F1-macro scores and F1 scores for each class in the following experiments.

5.2.1 Zero-Shot Inference

In this experiment, we evaluated the zero-shot inference performance of text-only language models and multimodal VLMs on selected datasets. The text-only models were the same language models used in the VLMs for text processing. The purpose of reporting the results on text-only models is to examine the necessity of image content for the fact-checking problem.

Table 5.1: Text-only and multimodal inference results

		MOCHEG				FACTIFY2			
Models	Inputs	Support	Refute	NEI	F1-macro	Support	Refute	NEI	F1-macro
Qwen-7B	text	0.533	0.262	0.169	0.321	0.524	0.458	0.281	0.421
Mistral-7B	text	0.505	0.281	0.216	0.334	0.575	0.561	0.093	0.409
Gemma-2b	text	0.610	0.462	0.315	0.462	0.562	0.119	0.083	0.255
Qwen-VL	text + image	0.168	0.472	0.186	0.275	0.463	0.460	0.369	0.431
Idefics2-8b	text + image	**0.619**	0.547	0.385	0.517	0.586	**0.644**	0.303	0.511
PaliGemma-3b	text + image	0.222	0.347	0.449	0.339	0.149	0.139	0.186	0.158
LVLM4FV	text	0.575	0.542	0.439	0.519	0.593	0.581	**0.560**	0.578
LVLM4FV	text + image	0.578	0.569	**0.457**	**0.535**	**0.678**	0.605	0.508	**0.597**
MOCHEG	text + image	0.490	**0.604**	0.282	0.459	0.547	0.621	0.275	0.481

Table 5.2: PaliGemma-3b fine-tuning results

		MOCHEG				FACTIFY2			
Models	Inputs	Support	Refute	NEI	F1-macro	Support	Refute	NEI	F1-macro
PaliGemma-3b	text + image	0.412	0.514	0.173	0.366	0.751	0.997	0.757	0.835

```
Assess the factuality of the following claim by
considering evidence. Only answer "supported",
"refuted" or "not enough info".
Claim: {claim}
Evidence: {evidence}
```

Figure 5.2: Prompt template

For the text-only models, the claim and evidence text were provided as a single prompt, as illustrated in Figure 5.2. Similarly, for each claim statement, the evidence text and evidence image were fed to the VLMs using a similar prompt template. Note that we reported results only for instances where the models responded with "supported," "refuted," or "not enough info." In other words, if the models did not provide a relevant justification, these cases were excluded from the reported results.

We also reported the performance of two baseline models, LVLM4V [125] and MOCHEG [55], for comparison. MOCHEG concatenates the claim, evidence and image to

generate CLIP [191] representations, employing attention mechanisms to update the claim representation based on the evidence. LVLM4V uses two-level prompting, formulating the problem as two binary questions and utilizing the Mistral [186] and LLaVa [192] models.

F1-macro scores along with F1 scores for each class are presented in Table 5.1 for both text-only and multimodal models. The results show that multimodality can enhance performance depending on the dataset and model configuration. For example, both Idefics-8b and LVLM4FV consistently outperformed their text-only counterparts, while Qwen-VL performed slightly better on the Factify2 dataset but worse on the Mocheg dataset. In contrast, PaliGemma consistently responded with, "sorry, as a base VLM I am not trained to answer this question" to test queries, suggesting that specific policies were implemented in the base VLM to prevent responses to ambiguous queries. As a result, PaliGemma's inference performance was significantly lower than that of its language model counterpart, Gemma-2b (see Appendix C.2 for response frequencies). The inference scores of Idefics2-8b suggest that images may provide additional information for fact-checking, likely due to its fine-tuning on a mixture of supervised and instruction datasets, which could explain its success on these datasets. Additionally, LVLM4V's prompting strategy appears more efficient, as it first checks whether the evidence is sufficient for verification before issuing a second prompt to verify or refute the claim.

Qualitative Analysis. A qualitative analysis was conducted to explore the types of claims that were correctly predicted by multimodal models but incorrectly predicted by text-only models. In this analysis, the predictions from both the text-only (Mistral-7B) and multimodal (Idefics2-8b) models were employed on the Mocheg dataset. Although for the fact-checking problem, textual contents are the primary source, images are shown to be useful. After examining the instances that are correctly predicted by the VLM but misclassified by the LLM, we found that such instances required image information to accurately verify the claims, as illustrated in Figure 5.6.

Fine-tuning PaliGemma-3b. Fact-checking requires long evidence with supporting images, making it computationally challenging to fine-tune the VLMs with mod-

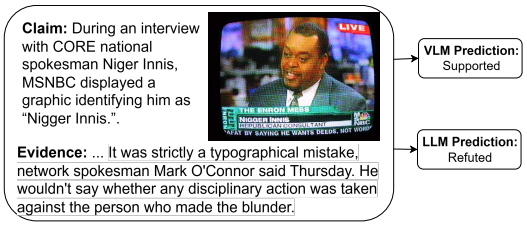


Figure 5.3: Supported claim

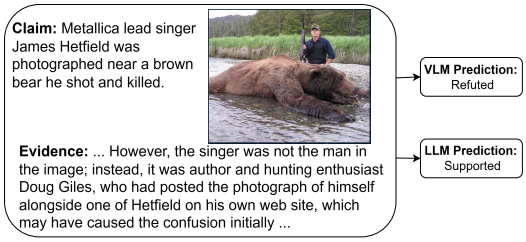


Figure 5.4: Refuted claim

Figure 5.5: Unproven claim

Figure 5.6: Qualitative examples for VLM and LLM inference predictions

erate batch sizes and sequence lengths on a single GPU. Therefore, we fine-tuned only the *PaliGemma-3b-pt-224* checkpoint using claim, evidence and claim image as input. The experimental details are given in Appendix C.3.

Evidence in the Mocheg dataset was collected from reference web articles. In contrast, Factify2 used the justifications provided by fact-checkers as evidence. As a result, Factify2's evidence is more concise and self-explanatory. However, models should interpret the knowledge from Mocheg's evidence sources to make a final decision. Because of the GPU memory considerations, evidence texts were cropped if they exceeded 768 words.

Fine-tuning results, presented in Table 5.2, show a significantly lower score of 0.366 on the Mocheg dataset compared to inference results, due to cropping of the evidence text. However, on the Factify2 dataset, the evidence texts were shorter and the model leveraged the key information for making a decision and achieved 0.835 F1-macro score. Note that, on the Factify2 challenge the best-performing model was Logically [193] which was also fine-tuned on Factify2 dataset and it achieved 0.897 F1-macro score. Due to computational constraints, we were unable to utilize the long text ev-

idence, particularly in the Mocheg dataset. As a result, we introduced a probing classifier instead of fine-tuning the selected VLMs.

Table 5.3: Intrinsic fusion of VLM embeddings: Feed-forward neural classification with VLM embeddings

		MOCHEG				FACTIFY2			
Model	Inputs	Support	Refute	NEI	F1-macro	Support	Refute	NEI	F1-macro
Qwen-VL	mm_claim	0.467	0.459	**0.463**	0.463	0.238	0.505	0.513	0.418
Idefics2-8b	mm_claim	**0.522**	0.535	0.399	0.485	0.427	0.516	0.471	0.471
PaliGemma-3b	mm_claim	0.495	0.510	0.451	0.485	0.398	0.387	0.503	0.429
Qwen-VL	mm_claim+mm_evd	0.483	0.561	0.417	0.487	**0.532**	0.443	0.469	0.481
Idefics2-8b	mm_claim+mm_evd	0.501	0.572	0.429	0.501	0.339	**0.674**	0.560	**0.524**
PaliGemma-3b	mm_claim+mm_evd	0.522	**0.592**	0.444	**0.519**	0.307	0.604	**0.575**	0.495

5.2.2 Intrinsic Fusion of VLM Embeddings

In this experiment, we examined whether inherently multimodal models effectively utilize both text and image information. First, we extracted embeddings from selected VLMs and fed these vector representations into a feed-forward multi-class classifier. We extracted the last hidden states and applied mean pooling to each token's embedding. In other words, the extracted embedding size was *(1, ntokens, ndim)*, where *ntokens* is the number of tokens and *ndim* is the dimension of each token embedding. Mean pooling provided a single embedding for each instance.

We provided two sets of inputs for extracting embeddings: *mm_claim* and *mm_evidence*. The *mm_claim* input consists of a claim and a corresponding image while the *mm_evidence* input consists of text evidence and an evidence image. For the second setting, we fed two input vectors to the classifier network: the *mm_claim* embedding and the *mm_evidence* embedding. This is because *mm_evidence* includes only the evidence representation - evidence image and evidence text - so we provided the claim information by feeding a second input to the classifier.

According to Table 5.3, the *mm_evidence* input setting improved F1-macro scores consistently for all models. This indicates that using both text and image evidence improved classification performance on both datasets. The results suggest that the selected VLMs effectively leverage information from evidence text and images on

both the Mocheg and Factify2 datasets.

5.2.3 Extrinsic Fusion of Language Model and Vision Encoder Embeddings

Table 5.4: Extrinsic fusion of embeddings: Feed-forward neural classification with distinct text and image embeddings

		MOCHEG				FACTIFY2			
Model	Inputs	Support	Refute	NEI	F1-macro	Support	Refute	NEI	F1-macro
Qwen-7B+Vit-bigG	claim+image	0.472	0.533	0.438	0.481	0.520	0.854	0.514	0.629
Mistral-7B+SigLIP	claim+image	0.515	0.555	**0.498**	0.522	0.095	0.951	**0.654**	0.566
Gemma-2b+SigLIP	claim+image	0.506	0.555	0.430	0.497	0.479	0.809	0.481	0.590
Qwen-7B+Vit-bigG	claim+clm_img+evd+evd_img	0.486	0.577	0.413	0.492	0.398	0.788	0.558	0.581
Mistral-7B+SigLIP	claim+clm_img+evd+evd_img	0.503	0.574	0.407	0.495	0.580	0.607	0.362	0.516
Gemma-2b+SigLIP	claim+clm_img+evd+evd_img	0.500	0.584	0.378	0.487	0.580	0.607	0.362	0.556
Qwen-VL	mm_claim+mm_image	0.528	0.515	0.462	0.502	0.318	0.806	0.642	0.589
Idefics2-8b	mm_claim+mm_image	**0.555**	0.578	0.452	**0.528**	0.437	**0.982**	0.593	**0.670**
PaliGemma-3b	mm_claim+mm_image	0.551	0.453	0.390	0.465	0.606	0.583	0.000	0.396
Qwen-VL	mm_text+mm_image	0.499	**0.612**	0.431	0.514	0.519	0.812	0.530	0.620
Idefics2-8b	mm_text+mm_image	0.526	0.541	0.458	0.509	0.319	0.825	0.547	0.564
PaliGemma-3b	mm_text+mm_image	0.467	0.512	0.447	0.475	**0.623**	0.681	0.001	0.435

Separate embeddings were extracted for text and image information from the vision encoders and language models, respectively. Afterward, we performed mean pooling to obtain one-dimensional vector representations for each instance. For this experiment, we had four input setups:

Input1 (claim+image): The claim representation was taken from the language model and the corresponding image representation was taken from the vision transformer.

Input2 (claim+claim_image+evd+evd_image): In addition to Input1, the evidence text representation was extracted from the language model and the evidence image representation was extracted from the vision transformer.

Input3 (mm_claim+mm_image): The embeddings extracted when the claim text is given to the VLM and the embeddings extracted when only the claim image is given were used separately.

Input4 (mm_text+mm_image): The embeddings extracted when all textual content is given to the VLM and the embeddings extracted when only the images are given

Table 5.5: Baseline classifiers' results

			MOCHEG				FACTIFY2			
Method	Model	Inputs	Support	Refute	NEI	F1-macro	Support	Refute	NEI	F1-macro
KNN	Qwen-VL	mm_claim	0.253	0.433	0.235	0.307	0.422	0.025	0.485	0.311
	Idefics2-8b	mm_claim	0.254	0.438	0.276	0.322	0.394	0.013	0.471	0.308
	PaliGemma-3b	mm_claim	0.237	0.435	0.250	0.307	0.410	0.009	0.471	0.293
	Qwen-VL	mm_claim+mm_evd	0.207	0.433	0.160	0.267	0.417	0.023	0.484	0.299
	Idefics2-8b	mm_claim+mm_evd	0.206	0.450	0.122	0.259	0.405	0.016	0.477	0.296
	PaliGemma-3b	mm_claim+mm_evd	0.150	0.457	0.148	0.252	0.401	0.017	0.471	0.296
SVM	Qwen-VL	mm_claim	0.375	0.453	0.273	0.367	0.234	0.156	0.512	0.301
	Idefics2-8b	mm_claim	**0.432**	0.491	**0.284**	**0.402**	0.268	**0.238**	0.479	0.217
	PaliGemma-3b	mm_claim	0.412	0.487	0.263	0.387	0.000	0.233	**0.533**	**0.328**
	Qwen-VL	mm_claim+mm_evd	0.380	0.490	0.233	0.368	0.583	0.046	0.023	0.320
	Idefics2-8b	mm_claim+mm_evd	0.392	0.514	0.231	0.379	**0.592**	0.187	0.181	0.255
	PaliGemma-3b	mm_claim+mm_evd	0.383	**0.521**	0.256	0.387	0.558	0.141	0.276	0.325

were used separately.

Inputs, except Input2, had two separate text and image embeddings. Only the second setup had four embeddings: claim embedding, claim image embedding, text embedding, and text image embedding. After extracting the embeddings, we trained the proposed probing classifier as described in Section 5.1.1 for multi-class veracity prediction. We extracted the embeddings for Input1 and Input2 using the selected multimodels' text and vision encoders that were also mentioned in Section 5.1.2.

According to Table 5.4, Idefics2 with the third input setup outperformed the other models on both datasets. Note that Idefics2 also performed better in zero-shot evaluations which could indicate that the model might have encountered similar data during pre-training. Therefore, it may leverage its pre-training knowledge while processing these claims.

5.2.4 Ablation Study

Our feed-forward classifier, illustrated in Figure 5.1, consists of two sequential linear layers. The first layer resizes each input tensor to a "hidden size" before concatenating the tensors. We chose this approach because there was a significant difference between the image and text embedding sizes. By reshaping each tensor to the same size before concatenation, we aimed to utilize both types of information more effec-

tively.

However, this approach has some limitations. If concatenation were performed before the first hidden layer, linear layers would be common for all models and input setups. In our approach, only the layers after concatenation are common so as the number of inputs increases, the number of learned parameters for the non-common layers also increases. Additionally, we did not validate the depth of the neural classifier and the network depth might be too shallow for the veracity detection task.

To assess whether the neural classifier effectively learns the intended task, we conducted an experiment using KNN and SVM classifiers with the same training embeddings as mentioned in Section 5.2.2. We set the number of neighbors (k), to seven which was decided after exploring consecutive values. Similarly, we trained SVM classifier with a linear kernel. As shown in Table 5.5, our approach outperformed the baselines on both datasets which implies that the proposed neural classifier leveraged the embeddings much better than the KNN and SVM classifiers on both datasets.

5.3 Discussion

First, we addressed RQ1 by conducting a zero-shot experiment to verify that multimodality improves performance depending on the dataset and model configuration, with models like Idefics-8b and LVLM4FV outperforming their text-only counterparts. Idefics2-8b benefits from image information while LVLM4V's efficient prompting strategy further enhances verification accuracy.

Additionally, the proposed intrinsic fusion pipeline which utilizes VLM embeddings, outperformed the VLMs' base inference performance (see Table 5.1 and Table 5.3). The only exception was the Idefics2 model on the Mocheg dataset, which had a 0.517 F1-macro inference score while the classifier achieved only a 0.501 F1-macro score. Since the probing classifier has only two layers, it might be too shallow for this dataset and model. Note that the primary goal of this chapter is not to achieve state-of-the-art scores for the selected datasets. Instead, we aim to evaluate whether recent VLMs improve performance on the fact-checking problem through multimodality or if fusing externally the information from distinct models achieves superior results.

Secondly, we addressed RQ2 by assessing how VLMs leverage text and image information. According to the results, for Idefics2-8b and Qwen-VL, multimodal embeddings were outperformed by discrete models (see Table 5.3 and Table 5.4). In other words, extracting separate embeddings resulted in higher F1-macro scores across all models. To be more specific, on the Mocheg dataset, the highest F1-macro scores for Qwen-VL and Idefics-8b were 0.514 and 0.528 respectively. Similarly, on the Factify2 dataset, the highest F1-macro scores were 0.629, 0.670 and 0.590 respectively. Although the best results were achieved with different input setups, for all of the best results, we extracted separate text and image embeddings. In contrast, when embeddings were extracted from inherently multimodal VLMs (as shown in Table 5.3), the maximum F1-macro scores were lower except PaliGemma-3b on Mocheg dataset. This indicates that for the given evaluation framework, using discrete text and image embeddings yielded higher F1-macro scores.

Besides, RQ3 was addressed by conducting an ablation study to examine how the proposed classifier leverages embeddings against KNN and SVM baselines. According to our evaluations, the proposed classifier utilized the extracted embeddings significantly better than the baseline approaches.

Finally, on the Mocheg dataset, the selected models struggle more on "not enough info" cases, as their lowest success rates, even in the best settings, were consistently associated with this class. This may be due to class relabeling, where the authors of the Mocheg dataset reannotated the "Mixture," "Unproven," and "Multiple" cases as "Not Enough Info" which may lead to confusion for the models. In contrast, on the Factify2 dataset, the trained classifier was more successful in distinguishing fake claims compared to other classes. This could be linked to the difference of data domains, as the genuine news was sourced from news agencies while fake claims were crawled from fact-checking sites and satirical articles.

CHAPTER 6

TEXT-BASED CAUSAL INFERENCE ON IRONY AND SARCASM DETECTION

Traditional NLP models can achieve accurate prediction results using statistical correlations within data. However, performance of the conventional methods mainly depends on the data distribution of the training and testing datasets. For this reason, analyzing causal relationships which utilize the data generating process are helpful to create robust models [20, 194]. More specifically, causal inference is a way of generating counterfactual explanations in hypothetical scenarios such as how the outcome variable is affected by an intervention on a treatment variable. The causal inference has been applied to create inferences on imaginary situations in several fields, but its practical applications in NLP have started to gain attention.

The cause-effect relationships of linguistic properties can be examined using causal inference by measuring the change in the outcome resulting from an intervention on a treatment. Under an imaginary scenario, the potential outcomes can be estimated by satisfying *ignorability*, *positivity*, and *consistency* assumptions (details given in Section 2.2). Usually, NLP applications rely on observational data, so randomly assigning texts is not feasible. In other words, to satisfy the ignorability assumption in observational studies while assigning treatment, there should not exist any unobserved confounders (predict both treatment and outcome). Identification is also another key aspect of causal inference for NLP, which suggests that the linguistic properties can be expressed using proxy labels [1, 195, 196]. Additionally, it is assumed that proxy labels can estimate the ground-truth causal relation of linguistic properties.

Many state-of-the-art NLP models can be considered black-box models, which receive text documents as input and generate an output dependent on the task. There-

fore, explaining and intervening in the predictions of such models remain a challenging problem [137, 24, 20]. Some studies examined the applicability of causal methods to interpret the black-box NLP models by generating counterfactual statements [23]. These works can be classified as data perspective [139] and model component perspective [197, 198] where the former is related to counterfactual statement generation and exploiting network artifacts is an example of the latter.

In this chapter, we focus on irony and sarcasm detection problem, and explore text-based causal inference by using the TextCause algorithm [1] to measure the causal effect of linguistic properties on this problem. The authors use DistilBERT [98] language model to adjust text and they are inspired by Veitch et al.'s CausalBERT study [132] which adapts BERT to adjust texts as a confounder. Additionally, they generate causal embeddings using causal topic models, which were adopted from Blei et al. [199].

Irony and sarcasm detection refers to way of verbal expressions such that the one's meaning is expressed through signifying just the opposite. Therefore, the problem includes difficulties and analyzing the causal relationships can provide insight for explainability of the generated models and improving the detection performance. The main contributions of this chapter can be summarized as follows:

- The causal effect of linguistic properties are examined in irony and sarcasm detection tasks using the TextCause algorithm.
- Latent confounders within text documents are modeled by using K-Means clustering and LDA topic modeling and their effects on the causal inference are analyzed.
- The obtained results provide insight in terms of the causal interpretability and explainability aspects.

6.1 Methods

In this work, we investigate the causal inference for irony and sarcasm detection problem, which involves text analysis. Therefore we apply text based causal inference

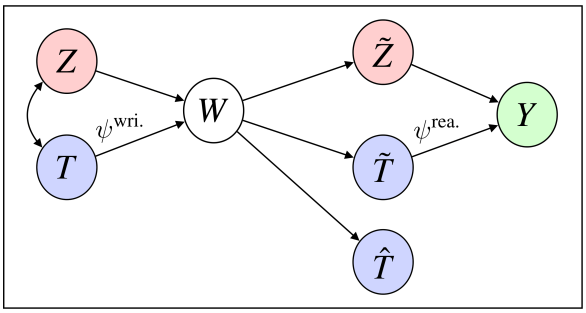


Figure 6.1: The structural causal model in Pryzant et al. [1]

algorithm, TextCause, [1]. In addition to adapting TextCause to irony/sarcasm detection problem, we extend the use of confounders by using unsupervised data analysis.

6.1.1 Text-based Causal Inference using TextCause

TextCause [1], employs the CausalBERT model [132] that adjusts text for causal inference. The key innovation of the TextCause algorithm is the assumption that neither the writer's intent nor the reader's perception can be identified from observational data. Therefore, the authors express the need to employ a proxy label $\hat{T}$ to estimate the causal effect of a linguistic property. In other words, they train a proxy classifier to capture both the writer's intent and the reader's perception. The proposed structural causal model is presented in Figure 6.1. According to this structural causal model, a writer writes a text W that contains a linguistic property T with other covariates Z. A reader's perception of that linguistic property is represented by $\widetilde{T}$ and $\widetilde{Z}$ and affects the outcome Y which can be estimated using a proxy label $\hat{T}$. Besides, the authors state that the bias due to proxy treatment decreases as the proxy classifier's accuracy increases. Therefore, for observational data, actual linguistic property T can be measured using proxy labels $\hat{T}$.

The conditional ignorability assumption of causal inference requires that the treatment assignment should be independent of the outcomes for observational data. In other words, this assumption states that we need to adjust for all confounders to estimate the causal effect of the treatment. The causal effect can be estimated using the Average Treatment Effect (ATE), which is formulated in ATE calculation Equation

(6.1).

$$ATE = E[Y; do(T = 1)] - E[Y; do(T = 0)] \tag{6.1}$$

ATE can be expressed as the difference between the interventional outcome (T=1) and the counterfactual outcome (T=0). However, text documents may contain some hidden confounders, such as tone and writing style, so we need to adjust the ATE for all confounders using Pearl's backdoor-adjustment [200]. Since the authors use proxy labels to estimate the ATE, the modified ATE estimation is given in Equation (6.2). The TextCause model uses DistilBERT to generate a representation of texts and employs the special classification token, CLS, to approximate the confounding information $\hat{Z}$. Therefore, the ATE estimator relies on the treatment, the language model representation of text and the one-hot encoding of the covariates. As a result, the model learns two vectors that corresponds to the language model representation and one-hot encoded covariates respectively.

$$ATE_{proxy} = E_W[E[Y|\hat{T} = 1, \widetilde{Z} = f(W)] - E[Y|\hat{T} = 0, \widetilde{Z} = f(W)]] \tag{6.2}$$

In addition to the text adjustment, another contribution of the TextCause algorithm is improving the recall of the proxy labels, which is motivated by lexicon induction [201] and label propagation [202]. The authors train logistic regression and pu-classifier models to predict proxy labels $\hat{T}^*$ and relabel the instances that labeled as $\hat{T}$=0 but predicted as $\hat{T}^*$=1. As a result, improved proxy labels and texts are required to measure the causal effect. Additional covariates and language model representation of a text should be adjusted as a confounder. Hence, the TextCause algorithm utilizes both proxy label improvement and text adjustment to estimate the causal effect of the desired linguistic property.

6.1.2 Unsupervised Data Analysis for Determining Confounders

While applying text-based causal inference on irony/sarcasm detection problem, the categories or groupings within the text collection is considered as a confounder. In

order to determine the subgroups, two different techniques[1], topic modeling and clustering, are used.

6.1.2.1 Topic Modeling

Topic modeling is a statistical method to discover latent topics in a corpus. It is an unsupervised technique that examines semantic structures in a text. Moreover, the topics represent a group of similar words that are determined by statistical models. A document can be a mixture of several topics with different proportions based on a word's appearance in one of the topics. Therefore, a document can be classified using topic modeling based on the words' relevance to the abstract topics.

Latent Dirichlet Allocation (LDA) [199] is one of the most popular topic modeling techniques. It is a generative statistical model that uses the Dirichlet priors for word-topic and document-topic distributions and represents documents as a mixture of topics where the distribution over words determines the proportions. Given a corpus with M documents where a document w_i contains N-words and α and β are the Dirichlet prior parameters, the probability distribution of a document can be expressed as in topic probability Equation (6.3). In this chapter, we lemmatized texts using SpaCy[2] and performed LDA to discover abstract topics that highlight several aspects of the document collection.

$$P(D, \alpha, \beta) = \prod_{m=1}^{M} \int P(\theta_m|\alpha)(\prod_{n=1}^{N} \sum_{Z_{mn}} P(Z_{mn}|\theta_m)P(W_{mn}|Z_{mn}, \beta))m\theta_m \quad (6.3)$$

6.1.2.2 Clustering

Texts are inherently high-dimensional, so a text should be encoded to a latent vector space. Sentence embeddings map sentences to vectors that can measure semantic similarity between sentences or text summarization. Transformers [203] made a remarkable impact on NLP tasks that passed previous models with a substantial margin.

[1] https://github.com/firatcekinel/Unsupervised-Data-Analysis
[2] https://spacy.io/

Reimers et al. [155] introduce S-BERT, which is a transformer-based sentence embedding model. S-BERT was built on top of the pre-trained BERT [19] model but uses siamese and triplet networks to extract semantically meaningful sentence embeddings. The S-BERT produces large-sized vectors as sentence embedding, which should be transformed into a lower-dimensional space for clustering. Dimensionality reduction techniques such as PCA [204], and t-SNE [205] can be applied to transform high-dimensional data into a lower-dimensional space by preserving the meaningful information in the data.

Clustering is an unsupervised machine learning technique that groups similar data instances together. K-Means clustering is one of the most popular clustering methods that assign n data points to k clusters where each data point is assigned to a cluster whose cluster center is the nearest. Since unsupervised models do not have a ground truth, metrics such as the silhouette coefficient can measure the clustering quality. We employ S-BERT to encode texts in a fixed-size latent space and applies dimensionality reduction using PCA or t-SNE. Finally, the transformed data is given to a K-Means model to group semantically similar texts.

6.1.3 Modeling Causal Inference for Irony and Sarcasm Detection

In this work, we explore the cause-effect relationship for irony and sarcasm detection on two scenarios. The treatments (T), outputs (Y) and confounders (Z) considered in the scenarios are as follows.

Case 1. We measure the effect of writing sarcastic posts (T) on the popularity of the post, number of likes, (Y) and consider subreddit category, cluster label (by the K-Means model) and the topic category (by the LDA model) as confounder (Z), separately.

Case 2. We examine whether putting an exclamation mark (!) affects irony detection. In other words, we explore whether the exclamation mark (T) affects the readers' perception of a text as ironic (Y). The cluster label and topic category were also

considered confounder (Z) in this scenario.

6.2 Experiments

6.2.1 Dataset and Settings

The first dataset that we use in our study is a Self-Annotated Reddit Corpus (SARC) [147] that contains 1.3 million sarcastic Reddit posts. It is a publicly-available dataset, and statements that end with "/s" marker, a common sarcastic marker of Reddit users, are annotated as sarcastic. Therefore, we can consider that the dataset might contain some false negative statements, such that there may be some statements that should be annotated as sarcastic but not marked as such. Moreover, we should not assume that all Reddit users know such markers, so the dataset might also contain some false positive statements. Secondly, we use a Turkish tweet dataset for irony detection [206, 145]. The dataset contains 300 non-ironic and 300 ironic tweets in Turkish, which were annotated manually.

The experiments are performed on Nvidia GeForce RTX 2080 Super GPU with 8GB memory. The computer also includes Intel i7-8700k CPU@3.7GHz with 12 cores. While implementing the model, Huggingface's multilingual DistilBERT [98] is used. It is a lighter BERT model that performs very close to the original model using significantly fewer parameters. Additionally, we performed some validation experiments to adjust hyperparameters such as epoch and learning rate. In Section 5.2, we present only the results with the best hyperparameter settings.

6.2.2 Results

6.2.2.1 Case 1 Results

In this experiment, we assume that the subreddit category, topic label and cluster label affect the treatment and outcome, so we consider these attributes as confounder.

Firstly, we gather the posts in "AskReddit" (Z=0), "news" (Z=1), "worldnews" (Z=2),

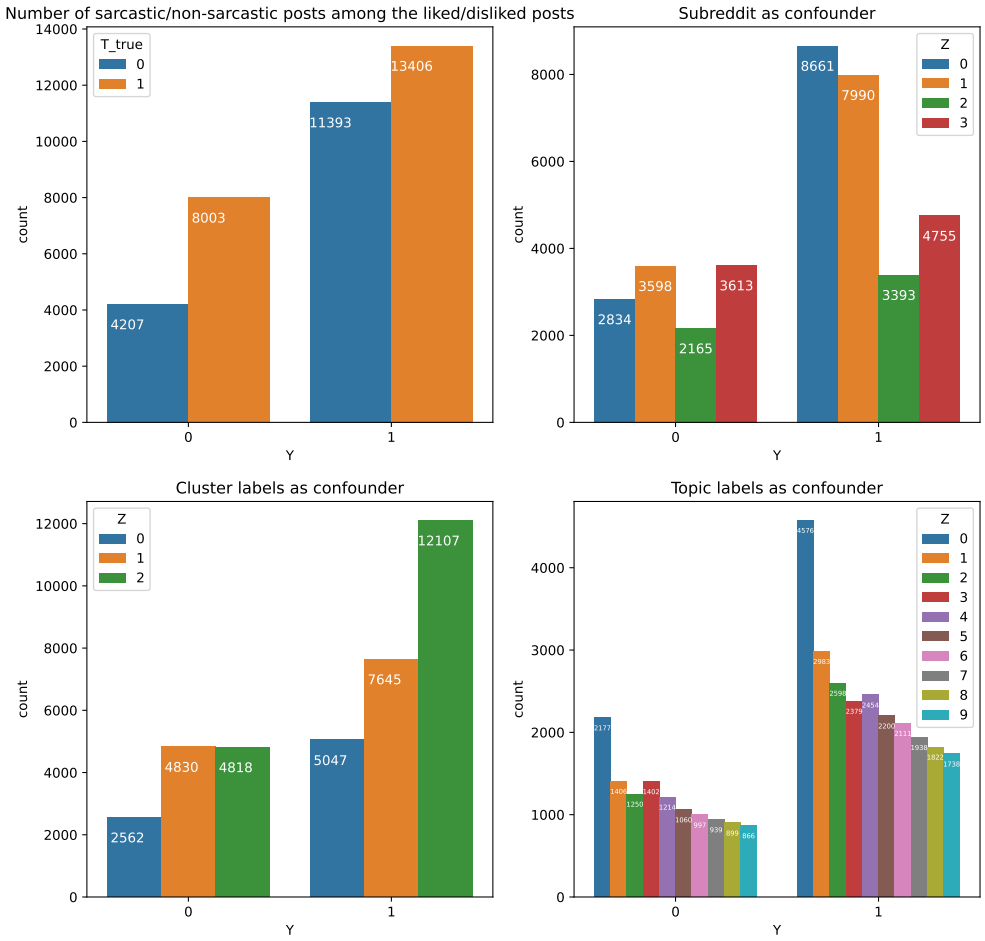
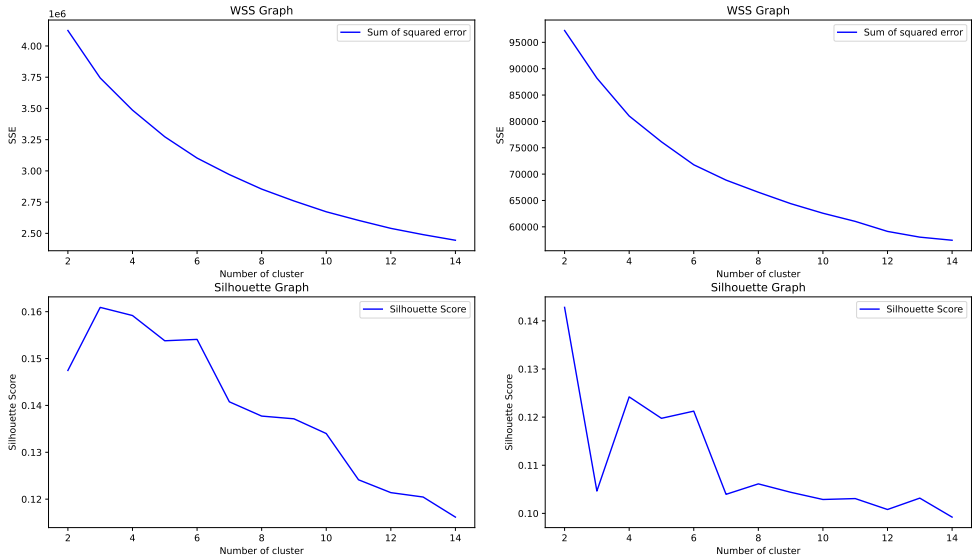


Figure 6.2: Number of Reddit posts for each confounder settings

and "politics" (Z=3) subreddits. If the posts' score is above five, we annotate them as "liked" comments. Besides if the posts' score is below 0, we annotate them as "disliked" comments. Overall, the number of comments satisfying these conditions are 37K approximately. ,The number of popular (liked) posts within each confounder is given in Figure 6.2.

Secondly, we assume that the LDA topic models could be used as a confounder. We measure the coherence score for various topic counts and observe that setting of 10 topics is a reasonable choice among a set of alternatives. The coherence score of this setting is 0.312. Likewise, we apply K-Means clustering to find optimal number of clusters with the collection of posts. According to Figure 6.3, K=3 is sensible among the selected set of values according to elbow analysis. Additionally, for K=3, PCA

and t-SNE plots are given in Figure 6.4.

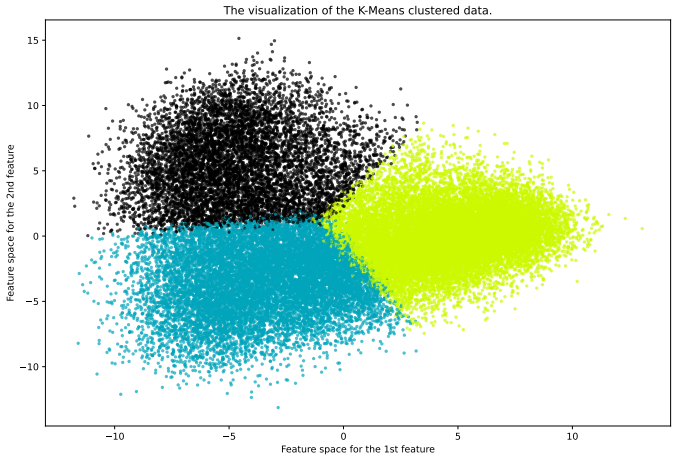


(a) Case 1 (b) Case 2

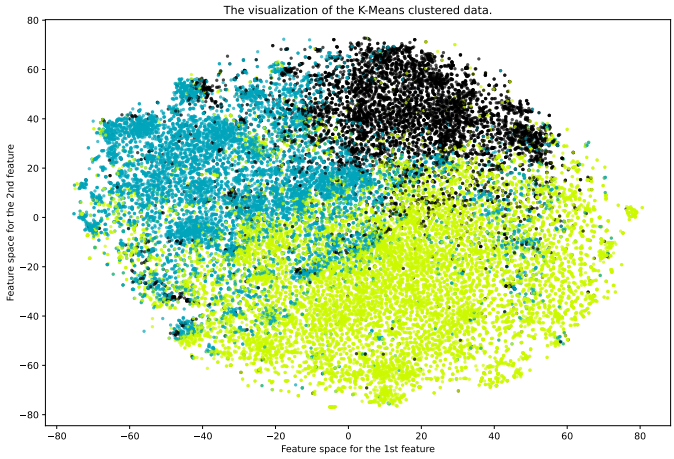
Figure 6.3: WSS and Silhouette Plots

Finally, we measure the ATE score using the subreddit category, topic label, and cluster label as a confounder. Since the TextCause model requires proxy labels, we trained a BERT model using 400K Reddit documents (80% - 20% train-val sets) from other categories. The accuracy of the proxy classifier on the selected subreddits is 78.6%, and the f1-score is also calculated as 0.806. The TextCause model measures the oracle ATE value using the ground truth sarcastic label. The unadjusted ATE measures the treatment effect without adjusting for any covariates. The T-boost values consider improved proxy treatments using pu classifier (to improve the recall for positive instances) and logistic regression. W adjust is another estimator that adjusts for text. Moreover, the last two estimates combine W adjust with T-boost.

We trained the TextCause algorithm for five epochs. According to the ATE scores in Table 6.1, adjusting for the topic label, cluster label, and subreddit category improves the ATE result. The oracle value suggests that the sarcastic writing style increases the chance of a post being liked between 6% and 10%. Additionally, the closest estimations are predicted by the T-boost reg model, and the TextCause models' subreddit and cluster label estimations are very close to the oracle estimator. However, when we



(a) PCA



(b) t-SNE

Figure 6.4: K-Means clusters of Reddit comments

adjust for topic labels, the unadjusted ATE estimator, which calculates ATE without adjusting for any covariate, becomes the second closest estimator overall.

6.2.2.2 Case 2 Results

In this experiment, we measure the effect of using an exclamation mark (!) on the irony. Since the treatment is evident, there is no need for a proxy label. We evaluate the causal question on the Turkish irony dataset, which is annotated by [145, 206]. As in the the first experiment, we consider the topic and cluster labels as a confounder.

Table 6.1: Case 1: Subreddit, topic and cluster labels were considered as confounder

Estimator	$ATE_{subreddit}$	ATE_{lda}	$ATE_{k-means}$
Oracle	0.0773	0.1029	0.0669
Unadjusted	0.1041	0.1041	0.1041
T-boost reg	**0.0742**	**0.1037**	**0.0639**
T-boost pu	0.0670	0.1005	0.0549
W adjust	0.0644	0.0659	0.0725
TextCause pu	0.0676	0.0776	0.0635
TextCause reg	0.0735	0.0719	0.0746

Figure 6.5 indicates the number of tweets for each confounder settings. According to the WSS and silhouette plots given in Figure 6.3, the highest silhouette score is measured when K=2. The clusters projected with the PCA and t-SNE are presented in Figure 6.6. On the other hand, for LDA model, 10 topics settings is a reasonable choice since the coherence score of this setting is measured as 0.7318.

We trained the TextCause algorithm for 15 epochs. According to the ATE results that are presented in Table 6.2, the treatment has a considerable impact on the posts' irony. However, contrary to our expectations, there is an inverse relationship between the treatment and the outcome. As seen in Figure 6.5, this is possibly due to that the number of ironic tweets that contain an exclamation mark is just 17% (51 out of 300 tweets) of the all ironic tweets. In addition, text adjustment for LDA topic labels estimates the closest prediction to the oracle value. However, for cluster labels the unadjusted setting was the closest among the all estimators. Note that, we do not present the results of the T-boost estimators because proxy labels were not appropriate in this setting.

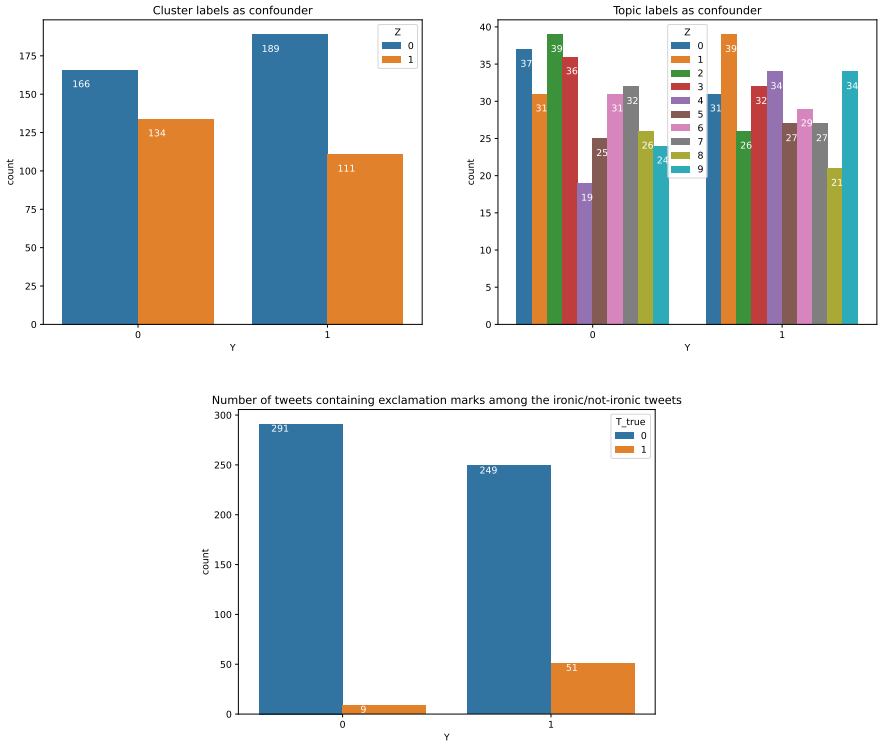
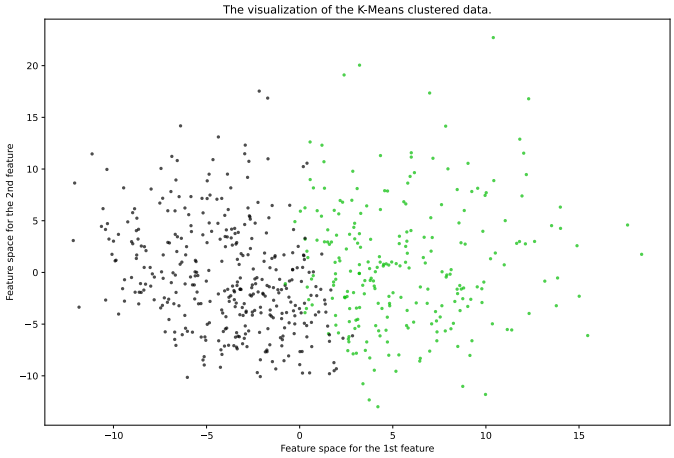


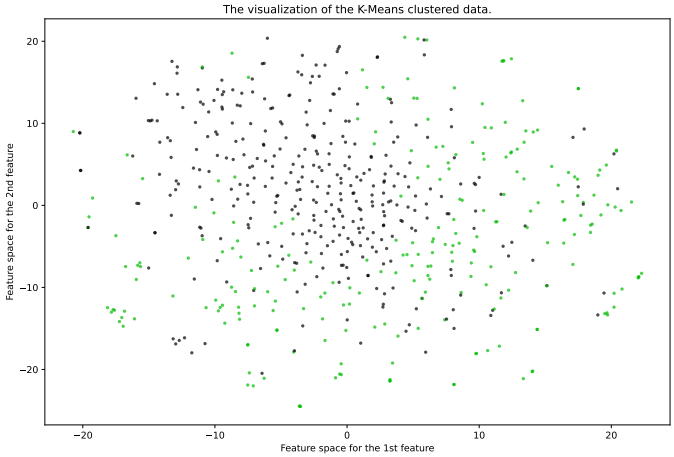
Figure 6.5: Number of tweets for each confounder settings

Table 6.2: Case2: Topic and cluster labels were considered as confounder

Estimator	$ATE_{k-means}$	ATE_{lda}
Oracle	-0.3955	-0.3451
Unadjusted	**-0.3889**	-0.3889
W adjust	-0.3506	**-0.3383**
TextCause pu	-0.0581	-0.0570
TextCause reg	-0.0292	-0.0204



(a) PCA



(b) t-SNE

Figure 6.6: K-Means clusters of tweets

CHAPTER 7

CONCLUSION

In this thesis, we evaluated the efficacy of multi-task training for veracity prediction and text summarization. We formulated text summarization as an explanation for the veracity prediction task, introducing a T5-based explainable multi-task fact-checking model. Our experimental results indicate that for the Flan-T5 model, joint training enhances text classification performance but slightly reduces summarization quality. In contrast, the T5 model shows a significant improvement in summarization results with minimal impact on classification performance.

We also introduced a novel Turkish fact-checking dataset, collected from three Turkish fact-checking sources, containing 3238 claims with accompanying evidence and summaries. Experiments demonstrated that fine-tuning a large language model on this dataset yields superior results compared to zero-shot and few-shot approaches, underscoring the value of datasets for languages with limited resources.

Additionally, we explored the use of VLMs for multimodal fact-checking. Our proposed pipeline extracts embeddings from the last hidden layer of selected VLMs, which are then processed by a simple feed-forward neural network for multi-class veracity classification. Initial zero-shot experiments confirmed the necessity of leveraging multimodal information for selected datasets, with the proposed pipeline outperforming base VLM inference performance. However, for all selected VLMs, multimodal embeddings were outperformed by discrete text-only and image-only models.

Lastly, we addressed the application of causal inference to text analysis. To be more specific, the TextCause algorithm [1] was employed to estimate the causal effect of sarcastic linguistic properties on a text's popularity, and use of punctuations, partic-

ularly (!) on understanding/detecting irony. Moreover, we performed unsupervised data analysis using clustering and topic modeling and utilized these methods' output for the causal inference. According to the measurements, cluster and topic labels may contain latent information on ironic linguistic properties and the popularity of the posts.

7.1 Limitations and Future Work

First, within the scope of the multi-task study, the T5 and Flan T5 models were pre-trained massively on English corpora. Consequently, the performance of these models on languages with limited resources may not be satisfactory. Secondly, the validation experiments revealed significant fluctuations in the model's performance when utilizing certain hyperparameter sets. Therefore, the hyperparameter optimization was a critical part of the evaluation process. Furthermore, the interpretability of the generated explanations may vary depending on the complexity of the text. Therefore, future research should address these limitations to enhance the robustness and applicability of our approach. Moreover, we aim to conduct a user study to evaluate the model's explanations' coherence and quality and also assess the explanations with the related studies. Moreover, transformer-based language models demand substantial computational and hardware resources.

Secondly, for the FCTR dataset that was released as a part of this thesis study, we did not process the collected data to ensure anonymization. The dataset encompasses fact-checked claims about public figures including politicians and artists. If any individual mentioned in a claim requests their removal, we can eliminate the associated claims. In addition, the data acquisition process adhered to the regulations of the Turkish text and data mining policy. This policy underlies that the datasets can be used exclusively for research purposes. Moreover, the Snopes dataset was collected in accordance with the Terms of Use set by Snopes. Therefore, anyone interested in accessing the Snopes dataset must send a request that includes a commitment to use the dataset only for non-commercial purposes.

Besides, as future work for multimodal fact-checking, we plan to employ VLMs as

assistants rather than as primary fact-checkers. To be more specific, the VLM can be used as an assistant that reviews the given text and image and returns a summary or justification to guide the text-only model for the fact-checking task. Since the LLMs are prone to hallucination and their accuracy depends on the quality of their training data which may be outdated or biased, incorporating knowledge grounding could be a more reliable strategy for real-world deployment. Note that we tested a limited number of models which may not fully capture the variability across different models and configurations. Additionally, the evaluations were performed on English datasets, restricting the assessment of multilingual capabilities. Furthermore, there is a potential risk that some dataset instances may overlap with the training data of the VLMs which could bias the evaluation results. Furthermore, while LLMs and VLMs are prone to hallucination, we did not perform any analysis on this phenomenon within the scope of this thesis study.

Finally, within the scope of causal analysis, the results can be reexamined in-depth in terms of explainability for future work. For instance, counterfactual statements that do not contain a specific linguistic property can be generated and fed into the causal-text model. The results can be examined in terms of invariance and sensitivity.

REFERENCES

[1] R. Pryzant, D. Card, D. Jurafsky, V. Veitch, and D. Sridhar, "Causal effects of linguistic properties," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4095–4109, 2021.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[3] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. Da San Martino, S. Shaar, H. Firooz, and P. Nakov, "A survey on multimodal disinformation detection," in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6625–6643, 2022.

[4] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, "Multimodal automated fact-checking: A survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5430–5448, 2023.

[5] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: a survey of deep learning techniques," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 101, 2023.

[6] S. Abdali, S. shaham, and B. Krishnamachari, "Multi-modal misinformation detection: Approaches, challenges and opportunities," 2024.

[7] R. F. Cekinel and P. Karagoz, " Explaining Veracity Predictions with Evidence Summarization: A Multi-Task Model Approach ," in *2024 IEEE International Conference on Big Data (BigData)*, (Los Alamitos, CA, USA), pp. 6924–6932, IEEE Computer Society, Dec. 2024.

[8] R. F. Cekinel, P. Karagoz, and Ç. Çöltekin, "Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in Turkish," in *Proceedings of the 2024 Joint International Conference on Computational Linguis-*

tics, Language Resources and Evaluation (LREC-COLING 2024) (N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, eds.), (Torino, Italia), pp. 4127–4142, ELRA and ICCL, May 2024.

[9] R. F. Cekinel, P. Karagoz, and Ç. Çöltekin, "Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies," in *Proceedings of the 31st International Conference on Computational Linguistics* (O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, eds.), (Abu Dhabi, UAE), pp. 4622–4633, Association for Computational Linguistics, Jan. 2025.

[10] R. F. Cekinel and P. Karagoz, "Text-based causal inference on irony and sarcasm detection," in *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 31–45, Springer, 2022.

[11] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[13] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[14] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6086–6093, 2020.

[15] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[16] T. Schuster, D. Shah, Y. J. S. Yeo, D. R. F. Ortiz, E. Santus, and R. Barzilay, "Towards debiasing fact verification models," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3419–3425, 2019.

[17] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[18] J. Pearl, "The do-calculus revisited," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 3–11, 2012.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, *et al.*, "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *arXiv preprint arXiv:2109.00725*, 2021.

[21] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

[22] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein, "Counterfactual invariance to spurious correlations: Why and how to pass stress tests," *arXiv preprint arXiv:2106.00545*, 2021.

[23] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, pp. 18–33, 2020.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[26] J. G. Richens, C. M. Lee, and S. Johri, "Improving the accuracy of medical diagnosis with causal machine learning," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[27] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances in neural information processing systems*, vol. 30, 2017.

[28] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022.

[29] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

[30] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, and M. Bansal, "HoVer: A dataset for many-hop fact extraction and claim verification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3441–3460, 2020.

[31] T. Schuster, A. Fisch, and R. Barzilay, "Get your vitamin C! robust fact verification with contrastive evidence," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, 2021.

[32] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, "The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task," in *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, (Dominican Republic), pp. 1–13, Association for Computational Linguistics, Nov. 2021.

[33] J. Kim, S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi, "FactKG: Fact verification via reasoning on knowledge graphs," in *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (Toronto, Canada), pp. 16190–16206, Association for Computational Linguistics, July 2023.

[34] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pp. 18–22, 2014.

[35] W. Y. Wang, ""Liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, 2017.

[36] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, "A richly annotated corpus for different tasks in automated fact-checking," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 493–503, 2019.

[37] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, and J. G. Simonsen, "Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697, 2019.

[38] K. Khan, R. Wang, and P. Poupart, "WatClaimCheck: A new dataset for claim entailment and inference," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (S. Muresan, P. Nakov, and A. Villavicencio, eds.), (Dublin, Ireland), pp. 1293–1304, Association for Computational Linguistics, May 2022.

[39] N. Kotonya and F. Toni, "Explainable automated fact-checking for public health claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, 2020.

[40] M. Sarrouti, A. B. Abacha, Y. M'rabet, and D. Demner-Fushman, "Evidence-based fact-checking of health-related claims," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3499–3512, 2021.

[41] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020.

[42] W. Zhang, Y. Deng, J. Ma, and W. Lam, "Answerfact: Fact checking in product question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2407–2417, 2020.

[43] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov, "Fake news detectors are biased against texts generated by large language models," *arXiv preprint arXiv:2309.08674*, 2023.

[44] G. K. Shahi and D. Nandini, "FakeCovid–a multilingual cross-domain fact check news dataset for COVID-19," *arXiv preprint arXiv:2006.11343*, 2020.

[45] J. Nørregaard and L. Derczynski, "DanFEVER: claim verification dataset for Danish," in *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, pp. 422–428, 2021.

[46] H. Ullrich, J. Drchal, M. Rỳpar, H. Vincourová, and V. Moravec, "Csfever and ctkfacts: acquiring czech data for fact verification," *Language Resources and Evaluation*, pp. 1–35, 2023.

[47] X. Hu, Z. Guo, G. Wu, A. Liu, L. Wen, and S. Y. Philip, "Chef: A pilot Chinese dataset for evidence-based fact-checking," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3362–3376, 2022.

[48] A. Barrón-Cedeno, T. Elsayed, R. Suwaileh, L. Màrquez, P. Atanasova, W. Zaghouani, S. Kyuchukov, G. Da San Martino, and P. Nakov, "Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 2: Factuality.," *CLEF (Working Notes)*, vol. 2125, 2018.

[49] A. Gupta and V. Srikumar, "X-Fact: A new benchmark dataset for multilingual fact checking," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 675–682, 2021.

[50] E. Raja, B. Soni, and S. K. Borgohain, "Fake news detection in Dravidian languages using transfer learning with adaptive finetuning," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106877, 2023.

[51] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto, "(Mis)information dissemination in WhatsApp: Gathering, analyzing and countermeasures," in *The World Wide Web Conference*, pp. 818–828, 2019.

[52] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6149–6157, 2020.

[53] G. Luo, T. Darrell, and A. Rohrbach, "NewsCLIPpings: Automatic generation of out-of-context multimodal media," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, 2021.

[54] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multimodal fact-checking of out-of-context images via online resources," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14940–14949, 2022.

[55] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, "End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743, 2023.

[56] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, *et al.*, "Factify 2: A multimodal fake news and satire news dataset," *arXiv preprint arXiv:2304.03897*, 2023.

[57] D. S. Nielsen and R. McConville, "MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3141–3153, 2022.

[58] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "Generating fact checking explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7352–7364, 2020.

[59] D. Stammbach and E. Ash, "e-fever: Explanations and summaries for automated fact checking," *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pp. 32–43, 2020.

[60] E. Brand, K. Roitero, M. Soprano, A. Rahimi, and G. Demartini, "A neural model to jointly predict and explain truthfulness of statements," *ACM Journal of Data and Information Quality*, vol. 15, no. 1, pp. 1–19, 2022.

[61] J. Vladika and F. Matthes, "Scientific fact-checking: A survey of resources and approaches," *arXiv preprint arXiv:2305.16859*, 2023.

[62] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, and J. Liu, "Content based fake news detection using knowledge graphs," in *International semantic web conference*, pp. 669–683, Springer, 2018.

[63] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 492–502, 2020.

[64] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[65] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3391–3401, 2018.

[66] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[67] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019.

[68] M. Hartmann, Y. Golovchenko, and I. Augenstein, "Mapping (dis-) information flow about the mh17 plane crash," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pp. 45–55, 2019.

[69] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[70] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, "Credibility-based fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media*, pp. 163–182, Springer, 2020.

[71] O. Ozcelik, C. Toraman, and F. Can, "Detecting misinformation on social media using community insights and contrastive learning," *ACM Trans. Intell. Syst. Technol.*, Dec. 2024.

[72] N. Kotonya and F. Toni, "Explainable automated fact-checking: A survey," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5430–5443, 2020.

[73] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 1003–1012, 2017.

[74] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32, 2018.

[75] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 395–405, 2019.

[76] Y.-J. Lu and C.-T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 505–514, 2020.

[77] A. Silva, Y. Han, L. Luo, S. Karunasekera, and C. Leckie, "Propagation2vec: Embedding partial propagation networks for explainable fake news early detection," *Information Processing & Management*, vol. 58, no. 5, p. 102618, 2021.

[78] M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "New explainability method for bert-based model in fake news detection," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[79] M. H. Gad-Elrab, D. Stepanova, J. Urbani, and G. Weikum, "Exfakt: A framework for explaining facts over knowledge graphs and text," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 87–95, 2019.

[80] N. Ahmadi, T.-T.-D. Truong, L.-H.-M. Dao, S. Ortona, and P. Papotti, "Rulehub: A public corpus of rules for knowledge graphs," *Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 4, pp. 1–22, 2020.

[81] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Evaluating adversarial attacks against multiple fact verification systems," Association for Computational Linguistics, 2019.

[82] P. Atanasova, D. Wright, and I. Augenstein, "Generating label cohesive and well-formed adversarial claims," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3168–3177, 2020.

[83] S.-C. Dai, Y.-L. Hsu, A. Xiong, and L.-W. Ku, "Ask to know more: Generating counterfactual explanations for fake claims," *arXiv preprint arXiv:2206.04869*, 2022.

[84] L. Cheng, R. Guo, K. Shu, and H. Liu, "Causal understanding of fake news dissemination on social media," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 148–157, 2021.

[85] W. Zhang, T. Zhong, C. Li, K. Zhang, and F. Zhou, "Causalrd: A causal view of rumor detection via eliminating popularity and conformity biases," in *IEEE*

INFOCOM 2022-IEEE Conference on Computer Communications, pp. 1369–1378, IEEE, 2022.

[86] Y. Li, K. Lee, N. Kordzadeh, and R. Guo, "What boosts fake news dissemination on social media? a causal inference view," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 234–246, Springer, 2023.

[87] W. Xu, Q. Liu, S. Wu, and L. Wang, "Counterfactual debiasing for fact verification," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Toronto, Canada), pp. 6777–6789, Association for Computational Linguistics, July 2023.

[88] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, and P. Nakov, "Fact-checking complex claims with program-guided reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Toronto, Canada), pp. 6981–7004, Association for Computational Linguistics, July 2023.

[89] H. Wang and K. Shu, "Explainable claim verification via knowledge-grounded reasoning with large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6288–6304, 2023.

[90] N. Ousidhoum, Z. Yuan, and A. Vlachos, "Varifocal question generation for fact-checking," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, (Abu Dhabi, United Arab Emirates), pp. 2532–2544, Association for Computational Linguistics, Dec. 2022.

[91] J. Yang, D. Vega-Oliveros, T. Seibt, and A. Rocha, "Explainable fact-checking through question answering," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8952–8956, IEEE, 2022.

[92] V. Schmitt, L.-F. Villa-Arenas, N. Feldhus, J. Meyer, R. P. Spang, and S. Möller, "The role of explainability in collaborative human-ai disinformation detection," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2157–2174, 2024.

[93] K. Kim, S. Lee, K.-H. Huang, H. P. Chan, M. Li, and H. Ji, "Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate," *arXiv preprint arXiv:2402.07401*, 2024.

[94] F. Zeng and W. Gao, "Justilm: Few-shot justification generation for explainable fact-checking of real-world claims," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 334–354, 2024.

[95] M. Leippold, S. Vaghefi, V. Muccione, J. Bingler, D. Stammbach, C. Colesanti Senni, J. Ni, T. Wekhof, T. Yu, T. Schimanski, *et al.*, "Automated fact-checking of climate change claims with large language models," *Available at SSRN 4731802*, 2024.

[96] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020.

[97] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[98] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[99] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3994–4003, 2016.

[100] A. Magooda, D. Litman, and M. Elaraby, "Exploring multitask learning for low-resource abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1652–1661, 2021.

[101] A. Kazemi, K. Garimella, D. Gaffney, and S. Hale, "Claim matching beyond English to scale global fact-checking," in *Proceedings of the 59th Annual Meet-*

ing of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4504–4517, 2021.

[102] A. Kazemi, Z. Li, V. Pérez-Rosas, S. A. Hale, and R. Mihalcea, "Matching tweets with applicable fact-checks across languages," *arXiv preprint arXiv:2202.07094*, 2022.

[103] O. Ozcelik, A. S. Yenicesu, O. Yildirim, D. S. Haliloglu, E. E. Eroglu, and F. Can, "Cross-lingual transfer learning for misinformation detection: Investigating performance across multiple languages," in *Proceedings of the 4th Conference on Language, Data and Knowledge* (S. Carvalho, A. F. Khan, A. O. Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch, and A. Salgado, eds.), (Vienna, Austria), pp. 549–558, NOVA CLUNL, Portugal, Sept. 2023.

[104] K.-H. Huang, C. Zhai, and H. Ji, "Concrete: Improving cross-lingual fact-checking with cross-lingual retrieval," in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1024–1035, 2022.

[105] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, and G. Da San Martino, "Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims," in *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, (Avignon, France), Springer, September 2018.

[106] M. Schlichtkrull, Z. Guo, and A. Vlachos, "Averitec: A dataset for real-world claim verification with evidence from the web," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 65128–65167, Curran Associates, Inc., 2023.

[107] E. Hoes, S. Altay, and J. Bermeo, "Leveraging chatgpt for efficient fact-checking," Apr 2023.

[108] C. Zhang, Z. Guo, and A. Vlachos, "Do we need language-specific fact-checking models? the case of chinese," *arXiv preprint arXiv:2401.15498*, 2024.

[109] T.-H. Cheung and K.-M. Lam, "Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 846–853, IEEE, 2023.

[110] Z. Yue, H. Zeng, Y. Zhang, L. Shang, and D. Wang, "Metaadapt: Domain adaptive few-shot misinformation detection via meta learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5223–5239, 2023.

[111] L. Tang, P. Laban, and G. Durrett, "Minicheck: Efficient fact-checking of llms on grounding documents," *arXiv preprint arXiv:2404.10774*, 2024.

[112] M. L. Bangerter, G. Fenza, D. Furno, M. Gallo, V. Loia, C. Stanzione, and I. You, "A hybrid framework integrating llm and anfis for explainable fact-checking," *IEEE Transactions on Fuzzy Systems*, 2024.

[113] R. Mediratta, J. Devasier, and C. Li, "Enabling automated fact checking of voting related claims using frame semantic parsing and semantic search," 2024.

[114] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13915–13916, 2020.

[115] C. Song, N. Ning, Y. Zhang, and B. Wu, "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks," *Information Processing & Management*, vol. 58, no. 1, p. 102437, 2021.

[116] W.-W. Du, H.-W. Wu, W.-Y. Wang, and W.-C. Peng, "Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification," 2023.

[117] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5696–5704, 2023.

[118] X. Gao, X. Wang, Z. Chen, W. Zhou, and S. C. Hoi, "Knowledge enhanced vision and language model for multi-modal fake news detection," *IEEE Transactions on Multimedia*, 2024.

[119] M. Liu, Y. Liu, R. Fu, Z. Wen, J. Tao, X. Liu, and G. Li, "Exploring the role of audio in multimodal misinformation detection," *arXiv preprint arXiv:2408.12558*, 2024.

[120] J. Wang, H. Zhang, C. Liu, and X. Yang, "Fake news detection via multi-scale semantic alignment and cross-modal attention," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2406–2410, 2024.

[121] J. Geng, Y. Kementchedjhieva, P. Nakov, and I. Gurevych, "Multimodal large language models to support real-world fact-checking," 2024.

[122] M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić, "Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models," *arXiv preprint arXiv:2404.12065*, 2024.

[123] L. Wang, X. Xu, L. Zhang, J. Lu, Y. Xu, H. Xu, and C. Zhang, "Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation," *arXiv preprint arXiv:2403.14171*, 2024.

[124] F. Yan, M. Zhang, B. Wei, K. Ren, and W. Jiang, "Sard: Fake news detection based on clip contrastive learning and multimodal semantic alignment," *Journal of King Saud University-Computer and Information Sciences*, p. 102160, 2024.

[125] S. Tahmasebi, E. Müller-Budack, and R. Ewerth, "Multimodal misinformation detection using large vision-language models," *arXiv preprint arXiv:2407.14321*, 2024.

[126] K. Keith, D. Jensen, and B. O'Connor, "Text and causal inference: A review of using text to remove confounding from causal estimates," in *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5332–5344, 2020.

[127] C. Fong and J. Grimmer, "Causal inference with latent treatments," *American Journal of Political Science*, 2019.

[128] C. Fong and J. Grimmer, "Discovery of treatments from text corpora," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1600–1609, 2016.

[129] Z. Wood-Doughty, I. Shpitser, and M. Dredze, "Challenges of using text classifiers for causal inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018, p. 4586, NIH Public Access, 2018.

[130] J. Yang, S. C. Han, and J. Poon, "A survey on extraction of causal relations from natural language text," *Knowledge and Information Systems*, pp. 1–26, 2022.

[131] J. Zhang, S. Mullainathan, and C. Danescu-Niculescu-Mizil, "Quantifying the causal effects of conversational tendencies," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–24, 2020.

[132] V. Veitch, D. Sridhar, and D. Blei, "Adapting text embeddings for causal inference," in *Conference on Uncertainty in Artificial Intelligence*, pp. 919–928, PMLR, 2020.

[133] N. Egami, C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart, "How to make causal inferences using texts," *arXiv preprint arXiv:1802.02163*, 2018.

[134] K. Keith, D. Rice, and B. O'Connor, "Text as causal mediators: Research design for causal estimates of differential treatment of social groups via language aspects," in *Proceedings of the First Workshop on Causal Inference and NLP*, pp. 21–32, 2021.

[135] D. Sridhar and L. Getoor, "Estimating causal effects of tone in online debates," in *International Joint Conference on Artificial Intelligence*, 2019.

[136] A. Koroleva, S. Kamath, and P. Paroubek, "Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations," *Journal of Biomedical Informatics*, vol. 100, p. 100058, 2019.

[137] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 447–459, 2020.

[138] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 606–615, 2016.

[139] A. Ross, T. Wu, H. Peng, M. E. Peters, and M. Gardner, "Tailor: Generating and perturbing text with semantic controls," *arXiv preprint arXiv:2107.07150*, 2021.

[140] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

[141] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, *et al.*, "Evaluating models' local decision boundaries via contrast sets," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

[142] A. Feder, N. Oved, U. Shalit, and R. Reichart, "Causalm: Causal model explanation through counterfactual language models," *Computational Linguistics*, vol. 47, no. 2, pp. 333–386, 2021.

[143] S. Ravfogel, G. Prasad, T. Linzen, and Y. Goldberg, "Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 194–209, 2021.

[144] E. B. Buyukbas, A. H. Dogan, A. U. Ozturk, and P. Karagoz, "Explainability

in irony detection," in *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 152–157, Springer, 2021.

[145] Y. Cemek, C. Cidecio, A. U. Öztürk, R. F. Çekinel, and P. Karagöz, "Investigating the neural models for irony detection on turkish informal texts," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2020.

[146] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, "Cascade: Contextual sarcasm detection in online discussion forums," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1837–1848, Association for Computational Linguistics, 2018.

[147] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[148] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[149] S. Chen, Y. Zhang, and Q. Yang, "Multi-task learning in natural language processing: An overview," *arXiv preprint arXiv:2109.09138*, 2021.

[150] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[151] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer.," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[152] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

[153] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pp. 150–157, 2003.

[154] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, 2004.

[155] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[156] D. Stammbach and G. Neumann, "Team domlin: Exploiting evidence enhancement for the fever shared task," in *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pp. 105–109, 2019.

[157] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[158] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[159] L. Stappen, F. Brunn, and B. Schuller, "Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL," *arXiv preprint arXiv:2004.13850*, 2020.

[160] H. Lin, P. Yi, J. Ma, H. Jiang, Z. Luo, S. Shi, and R. Liu, "Zero-shot rumor detection with propagation structure via prompt learning," AAAI'23/IAAI'23/EAAI'23, AAAI Press, 2023.

[161] G. Glavaš, M. Karan, and I. Vulić, "Xhate-999: Analyzing and detecting abusive language across domains and languages," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6350–6365, 2020.

[162] S. Haider, L. Luceri, A. Deb, A. Badawy, N. Peng, and E. Ferrara, "Detecting

social media manipulation in low-resource languages," in *Companion Proceedings of the ACM Web Conference 2023*, pp. 1358–1364, 2023.

[163] J. Du, Y. Dou, C. Xia, L. Cui, J. Ma, and S. Y. Philip, "Cross-lingual COVID-19 fake news detection," in *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 859–862, IEEE, 2021.

[164] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *International Conference on Learning Representations*, 2019.

[165] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[166] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[167] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023.

[168] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.

[169] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[170] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[171] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

and Short Papers), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[172] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford Alpaca: An instruction-following LLaMA model." https://github.com/tatsu-lab/stanford_alpaca, 2023.

[173] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.

[174] J. Tiedemann and S. Thottingal, "OPUS-MT — Building open translation services for the World," in *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, (Lisbon, Portugal), 2020.

[175] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

[176] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040, 2022.

[177] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[178] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "Dora: Weight-decomposed low-rank adaptation," *arXiv preprint arXiv:2402.09353*, 2024.

[179] J. Kunz and M. Kuhlmann, "Classifier probes may just learn from linear context features," in *Proceedings of the 28th International Conference on Computational Linguistics* (D. Scott, N. Bel, and C. Zong, eds.), (Barcelona, Spain (Online)), pp. 5136–5146, International Committee on Computational Linguistics, Dec. 2020.

[180] B. Adelmann, W. Menzel, and H. Zinsmeister, "The impact of word embeddings on neural dependency parsing," in *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pp. 1–13, 2021.

[181] J. Kunz and M. Kuhlmann, "Test harder than you train: Probing with extrapolation splits," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 15–25, 2021.

[182] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[183] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[184] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021.

[185] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?," 2024.

[186] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[187] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[188] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[189] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[190] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, and S. Kumar, "Factify 2: A multimodal fake news and satire news dataset," 2023.

[191] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[192] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[193] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, and A. Bandhakavi, "Logically at factify 2022: Multimodal fact verification," *arXiv preprint arXiv:2112.09253*, 2021.

[194] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[195] L. Lucy, D. Demszky, P. Bromley, and D. Jurafsky, "Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks," *AERA Open*, vol. 6, no. 3, p. 2332858420940312, 2020.

[196] R. Voigt, N. P. Camp, V. Prabhakaran, W. L. Hamilton, R. C. Hetey, C. M. Griffiths, D. Jurgens, D. Jurafsky, and J. L. Eberhardt, "Language from police body camera footage shows racial disparities in officer respect," *Proceedings of the National Academy of Sciences*, vol. 114, no. 25, pp. 6521–6526, 2017.

[197] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," *arXiv preprint arXiv:1802.00541*, 2018.

[198] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," *arXiv preprint arXiv:1811.04376*, 2018.

[199] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[200] J. Pearl, *Causality*. Cambridge university press, 2009.

[201] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, vol. 2016, p. 595, NIH Public Access, 2016.

[202] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," 2002.

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

[204] I. Jolliffe, "Principal component analysis springer-verlag," *doi*, vol. 10, p. b98835, 2002.

[205] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[206] A. U. Ozturk, Y. Cemek, and P. Karagoz, "Ironytr: Irony detection in turkish informal texts," *International Journal of Intelligent Information Technologies (IJIIT)*, vol. 17, no. 4, pp. 1–18, 2021.

[207] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, pp. 759–766, 2017.

[208] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, pp. 99–114, 1949.

[209] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

APPENDIX A

APPENDIX A

A.1 More Examples for Generated Summaries

EXAMPLE 1

Claim: “While California is dying . . . Gavin (Newsom) is vacationing in Stevensville, MT!”

Evidence: ... A Facebook post said, "While California is dying . . . Gavin (Newsom) is vacationing in Stevensville, MT!" There is no evidence of this. Newsom’s office said he has not been vacationing in Stevensville, and so did the lieutenant governor’s office, the Montana governor’s office and the Stevensville mayor. Newsom gave a live press conference from a California restaurant on May 18, the day the post went up.

Gold Summary: Newsom has not been vacationing in Stevensville in recent weeks, his office said. The Montana governor’s office and Stevensville mayor said the same thing.

Generated Summary: A Facebook post said Newsom was vacationing in Stevensville, Montana. The governor has not been there in recent weeks.

Gold Label: FALSE

Predicted Label: FALSE

EXAMPLE 2

Claim: Treating at the Earliest Sign of MS May Offer Long-Term Benefit

Evidence: ... this particular study is adding to evidence and is not particularly novel. “Our study adds to the evidence supporting treatment at the earliest sign of the dis-

ease and indicates that early treatment has a long-lasting effect on disease activity," a researcher says in the release. However, another novelty of the study is starting treatment after one event and not waiting for any further events. This could lead to a much larger market for the drug. It could be concerning since some patients only have a few recurrences in their lifetime. There was no unjustifiable language.

Gold Summary: This release summarizes a study that followed 11-year treatment outcomes of two groups of patients with early symptoms of multiple sclerosis (MS). All of the patients had experienced a first episode suggestive of MS, called clinically isolated syndrome (CIS), and a magnetic resonance imaging (MRI) scan suggesting MS. Up to 85 percent of people with CIS go on to develop MS. The two patient groups were randomized to either receive interferon beta-1 by injection immediately or receive a placebo for two years. After two years, the placebo group could receive their treatment of choice. The patients who received the drug immediately appeared to have fewer relapses over the 11 years of study, according to the lead author, but no advantages were observed in other patient outcomes such as overall disability and MRI scans. The release would have been stronger if it provided more context for judging the reduced number of relapses, especially in light of the two groups showing no difference in disability. The release also omitted mention of significant potential conflicts of interest, the high annual cost of interferon beta-1 and the drug's side effects. Society faces tough questions because of the high prices of drugs that target the immune system, such as interferon beta-1b for multiple sclerosis. This news release suggests drug therapy should begin sooner for patients without including any of the controversy on cost playing out in the United States and Europe. Last year authors of a different research study, also published in Neurology, called the rise in MS drug prices "alarming" and wrote it was also puzzling, since some of the drugs are not new. "Why the costs of MS DMTs [disease modifying therapies] in the United States have risen so dramatically is uncertain. However, the simplest explanation is that pharmaceutical companies raise prices of new and old MS DMTs in the United States to increase profits and our health care system puts no limits on these increases,' they wrote. Why is cost so important to mention? This drug study, in particular, has significant potential conflicts of interest, discussed below under the Funding Sources & Conflict of Interest Disclosure criteria.

Generated Summary: This news release describes a study that found early treat-

ment with interferon beta-1b (brand name Betaseron) for patients with demyelination, often an early sign of multiple sclerosis (MS), was associated with a lower rate of relapses at 11 years. The release does a good job of explaining the study and its limitations. However, it did not mention any potential harms of treatment. There are significant side effects reported to therapy with interferon beta-1b, including depression and thoughts of suicide, according to the National Library of Medicine. No harms were mentioned in the study. We also wonder — if patients routinely receive a therapy before official diagnosis — whether that risks treating people for a condition they do not eventually develop. The published study states that about 85% of people who encounter a single episode of demyelination (often an early sign of MS) go on to develop MS.

Gold Label: MIXTURE

Predicted Label: MIXTURE

EXAMPLE 3

Claim: Queen Elizabeth II wore a Burmese Ruby Tiara as a slight against U.S. President Donald Trump.

Evidence: ... Given that Queen Elizabeth has worn this tiara in the company of a variety of world leaders and for a number of different special occasions, those claiming that she chose to wear it in June 2019 specifically as a slight against President Trump appear to be revealing more about their own political beliefs than the queen's.

Gold Summary: What's true: Queen Elizabeth II wore a Burmese Ruby Tiara while meeting President Donald Trump in June 2019. The 96 rubies that adorn this tiara are said to symbolically protect the wearer from 96 diseases. What's false: No evidence exists that the queen specifically chose this tiara as a slight against Trump, and the queen has worn this same tiara on several other occasions and in the company of a wide range of world leaders.

Generated Summary: Did Queen Elizabeth II Wear a Burmese Ruby Tiara as a Sighting of Disrespect?

Gold Label: FALSE

Predicted Label: UNPROVEN

Table A.1: Grid search of loss coefficients

Veracity (a), Summary (b) loss coefficients	Veracity label coefficients	Rouge-1	Rouge-2	Rouge-L	F1-macro	F1-weighted
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=5	31,99	14,14	28,18	51,14	66,66
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=7	31,93	14,26	28,46	**60,76**	**73,16**
a=0.7, b=0.3	mixture_coeff=1.75, unproven_coeff=9	31,87	14,16	28,13	48,62	63,93
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=5	31,25	13,81	27,59	54,71	69,92
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=7	32,36	**14,59**	28,67	57,22	71,47
a=0.6, b=0.4	mixture_coeff=1.75, unproven_coeff=9	31,85	14,21	28,16	54,14	68,48
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=5	32,52	14,50	**28,74**	56,71	69,86
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=7	31,87	13,94	27,09	52,00	67,25
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=9	31,71	13,88	28,19	51,12	65,78
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=5	31,02	13,50	27,53	50,94	65,48
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	31,82	14,00	28,12	55,87	68,57
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	31,96	14,42	28,40	56,52	69,93
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=5	31,43	14,03	27,75	50,59	65,76
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	31,96	14,38	28,28	55,62	68,57
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	**32,54**	14,48	28,69	60,07	72,50
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=5	31,78	13,86	28,04	58,73	72,20
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=5	32,27	14,32	28,64	58,02	72,19
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=7	31,05	13,44	27,49	50,96	65,48
a=0.8, b=0.2	mixture_coeff=1.75, unproven_coeff=9	32,03	13,74	28,06	57,59	70,41
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=5	32,00	14,29	28,38	56,31	70,19
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=7	31,82	14,16	28,14	55,05	69,52
a=0.5, b=0.5	mixture_coeff=1.75, unproven_coeff=9	31,87	14,15	28,22	58,33	72,34
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=5	32,42	14,11	28,50	54,14	67,34
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=7	32,03	14,20	28,31	58,87	71,89
a=0.8, b=0.2	mixture_coeff=2.5, unproven_coeff=9	31,84	13,93	28,07	58,10	71,95
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=5	31,85	14,25	28,13	52,58	66,45
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	32,33	14,18	28,48	60,33	73,11
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	31,90	14,14	28,27	55,56	70,32

A.2 Grid Search of Static Loss Coefficients

We performed an ablation study to explore candidate values to find an optimal set of hyper-parameters for our multi-task model. We performed a grid search using PUBHEALTH [39] dataset to determine the optimal set of loss coefficients. The experimental results are presented in Table A.1. Note that, we kept the linear layers' size (for veracity prediction), dropout probability, batch size and number of epoch constant.

Table A.2: Grid search of hidden layer size

Veracity (a), Summary (b) loss coefficients	Veracity label coefficients	Hidden Dim	Rouge-1	Rouge-2	Rouge-L	F1-macro	F1-weighted
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	16	31,82	14,00	28,12	55,87	68,57
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	16	31,96	14,42	28,40	56,52	69,93
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	16	31,96	14,38	28,28	55,62	68,57
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	16	32,54	14,48	28,69	60,07	72,50
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	16	32,33	14,18	28,48	60,33	**73,11**
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	16	31,90	14,14	28,27	55,56	70,32
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	32	31,97	14,21	28,23	51,14	65,77
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	32	31,83	14,00	28,05	57,25	68,34
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	32	31,82	14,21	28,14	58,96	60,78
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	32	32,08	14,09	28,34	52,47	65,67
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	32	32,07	14,33	28,32	59,18	71,91
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	32	31,79	14,13	28,29	49,99	61,82
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	64	32,55	14,54	28,60	**60,93**	72,51
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	64	**32,69**	**14,71**	**28,84**	49,08	62,63
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	64	31,97	14,28	28,30	44,73	57,52
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	64	31,98	14,19	28,33	57,78	72,52
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	64	31,78	13,95	28,01	59,22	72,20
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	64	31,63	13,99	27,89	53,21	66,03
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	128	31,97	14,21	28,23	51,14	65,77
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	128	31,83	14,00	28,05	57,25	68,34
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	128	31,82	14,21	28,14	48,42	60,78
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=9	128	32,08	14,09	28,34	52,47	65,67
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=7	128	31,79	14,13	28,29	49,99	61,82
a=0.5, b=0.5	mixture_coeff=2.5, unproven_coeff=9	128	32,55	14,54	28,60	**60,93**	72,51
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=7	256	32,07	14,33	28,32	59,18	71,91
a=0.7, b=0.3	mixture_coeff=2.5, unproven_coeff=9	256	**32,69**	**14,71**	**28,84**	49,08	62,63
a=0.6, b=0.4	mixture_coeff=2.5, unproven_coeff=7	256	31,97	14,28	28,30	44,73	57,52

A.3 Grid Search of Hidden Layer Dimensions for Veracity Prediction

We also performed another ablation study to discover the optimal hidden layer size of the classification head of our multi-task model using the PUBHEALTH [39] dataset. The experimental results are presented in Table A.2. Note that, we kept the dropout probability, batch size and number of epochs constant.

APPENDIX B

APPENDIX B

B.1 Topic Modeling

Table B.1: Topic distribution in the FCTR dataset

Dataset	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
FCTR500-train	39	64	105	49	116	27
FCTR500-val	8	10	10	9	9	4
FCTR500-test	6	9	7	9	15	4
FCTR1000-train	73	132	174	130	237	54
FCTR1000-val	9	16	20	18	29	8
FCTR1000-test	12	11	19	21	35	2
FCTR	293	472	524	600	927	167

Table B.2: Topic distribution in the Snopes dataset

Dataset	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Snopes-train	206	1063	386	260	553	327	193
Snopes-val	26	125	52	27	73	48	23
Snopes-test	25	124	43	29	75	50	27

Topic modeling is a method for discovering abstract topics in a collection of documents. Latent topics indicate the patterns in the data that can be inferred by the relationships between words that occur in the documents. The output of a topic mod-

Table B.3: Representative words in *FCTR* dataset

Topics	Representative Words (transl.)
Topic1	claim, news, person, sharing, information, account, share, be, child, use
Topic2	photograph, image, account, sharing, share, claim, video, name, view, use
Topic3	country, Turkiye, year, history, claim, data, take, be, state, Turkic
Topic4	vaccine, be, virus, claim, work, human, disease, research, person, impact
Topic5	video, claim, news, be, statement, sharing, name, history, eat, talk
Topic6	use, product, breeding, water, electricity, plane, production, year, logo, claim

eling is a set of abstract topics that are represented by a list of the most representative words in the topic. In our analysis, Latent Dirichlet Allocation (LDA) [199] topic modeling is applied to the *Snopes* and *FCTR* datasets to explore the latent patterns using the coherence metric. The coherence score can be used to evaluate the semantic similarity between the words in a topic.

The topic distributions for each data split are given in Table B.1 and Table B.2 respectively. Even though we did not split the datasets according to the topic ratios, the most dominant and the least frequent topics were preserved in all data splits. For instance, in the *FCTR* dataset, The fifth topic is the most frequent topic in all subsets except *FCTR500-val* in which the given topic is not the most dominant topic by a small margin. Additionally, the sixth topic is the least frequent topic in all splits.

We utilized lemmatization, employing the Spacy library for English [1] and the Zeyrek library for Turkish [2]. Table B.3 and Table B.4 display the most representative words for each topic. The coherence score for the Turkish dataset within these topics was

[1] https://spacy.io/models/en
[2] https://zeyrek.readthedocs.io/en/latest/

Table B.4: Representative words in *Snopes* dataset

Topics	Representative Words
Topic1	animal, water, world, report, military, human, fire, Russian, area, Russia
Topic2	say, people, year, man, know, take, make, time, go, get
Topic3	image, photograph, show, video, picture, take, create, appear, film, real
Topic4	Trump, president, Obama, White House, former, Clinton, President Donald, tweet, Donald Trump, say
Topic5	post, article, news, Facebook, claim, story, publish, report, page, com
Topic6	state, law, government, report, vote, bill, United States, federal, election, claim
Topic7	covid, vaccine, health, study, drug, medical, cause, disease, use, patient

0.388, and the perplexity score was -7.699. The average entropy value per document was calculated as 1.50, suggesting a moderate topic distribution level. Similarly, the Snopes dataset achieved a coherence score of 0.450 and a perplexity score of -8.796. Moreover, the average entropy score per document was found to be 1.94 which might indicate that the documents cover multiple related topics without a strong focus on a single one.

B.2 NELA Features

News Landscape (NELA) features [207] are manually crafted content-based textual attributes for news veracity detection. The authors divided the features into six classes: style, complexity, bias, affect, moral and event. We applied NELA features to examine the discrepancies of the features for fake and true claims in the FCTR dataset and conducted Tukey's pairwise test [208] to identify statistically significant

Table B.5: Statistically significantly different NELA features

Subset	Feature name	Adjusted p-value
FCTR500	allcaps	0.023
FCTR500	avg_wordlen	0.018
FCTR500	coleman_liau_index	0.018
FCTR500	lix	0.032
FCTR1000	NNP	0.049
FCTR1000	avg_wordlen	0.048
FCTR1000	coleman_liau_index	0.045
FCTR1000	lix	0.048

differences.

Table B.5 presents features that exhibit statistically significant distinctions for *FCTR500* and *FCTR1000*. We computed the NELA features for only claim statements and the results indicate that only a few features demonstrate significant divergence for fake and true claims.

APPENDIX C

APPENDIX C

C.1 Hyperparameter Values for the Best Models

We set the number of epochs to 20, enabling early stopping with the patience of 5 and monitoring the validation loss. We used the Adam optimizer in combination with a cosine scheduler, employing a warm-up ratio of 0.05. Moreover, we adjusted the cross-entropy loss weight of the neural network according to the inverse class ratios. In this way, the classifier was penalized more for the misclassifications of the minority classes.

We performed a grid search to explore the following parameter space for the results given in Table 5.3 and Table 5.4:

learning rate: { 0.00001, 0.0001, 0.001, 0.01, 0.1}, ,
batch size: {32, 64, 128},
hidden size (h in Figure 5.1): {128, 256, 512 } and
dropout: {0.05, 0.1, 0.2, 0.4}.

The parameter settings for the best results are detailed in Table C.1.

C.2 Zero-shot Model Response Frequencies

We used the prompt template shown in Figure 4.3 for all models in the zero-shot inference experiments. We expected the models' responses to contain either "supported," "refuted," or "not enough info." If a model's response did not contain these labels, we ignored those instances. Additionally, we observed that PaliGemma consistently re-

		MOCHEG				FACTIFY2			
Embedding	Input	Batch	LR	Hidden size	Dropout	Batch	LR	Hidden size	Dropout
Qwen-VL	claim	32	0.01	128	0.1	64	0.001	128	0.1
idefics2-8b	claim	32	0.01	256	0.05	32	0.0001	128	0.1
PaliGemma-3b	claim	32	0.01	512	0.05	64	0.0001	128	0.05
Qwen-VL	claim+evd	64	0.01	256	0.05	32	1E-05	256	0.05
idefics2-8b	claim+evd	64	0.01	512	0.1	32	0.001	256	0.1
PaliGemma-3b	claim+evd	64	0.001	256	0.1	64	1E-05	512	0.1
Qwen-7B+Vit-bigG	input1	128	0.01	512	0.1	32	0.001	128	0.1
Mistral-7B+SigLIP	input1	64	0.001	512	0.1	128	0.001	256	0.2
Gemma-2b+SigLIP	input1	64	0.01	512	0.1	128	0.001	128	0.1
Qwen-7B+Vit-bigG	input2	32	0.001	256	0.4	64	0.001	128	0.1
Mistral-7B+SigLIP	input2	64	0.01	512	0.1	64	0.001	256	0.4
Gemma-2b+SigLIP	input2	64	0.001	512	0.2	64	0.001	256	0.4
Qwen-VL	input3	32	0.001	512	0.2	128	0.001	512	0.1
Idefics2-8b	input3	128	0.001	512	0.1	128	0.01	512	0.1
PaliGemma-3b	input3	64	0.001	256	0.1	64	0.001	256	0.2
Qwen-VL	input4	64	0.001	512	0.1	128	0.001	128	0.4
Idefics2-8b	input4	128	0.001	128	0.4	128	0.001	128	0.1
PaliGemma-3b	input4	64	0.001	256	0.2	32	0.001	512	0.4

Table C.1: Parameter settings for the best models

Model	Mocheg (1655)	Factify2 (7273)
Qwen-7B	1366 (82.5%)	4335 (59.6%)
Mistral-7B	1361 (82.2%)	5756 (79.1%)
Gemma-2B	1617 (97.7%)	6136 (84.4%)
Qwen-VL	1646 (99.5%)	6483 (89.1%)
Idefics2-8b	1653 (99.9%)	5873 (80.7%)
PaliGemma-3b	320 (19.3%)	91 (1.2%)

Table C.2: Zero-shot response frequencies

sponded with "sorry, as a base VLM I am not trained to answer this question," which could be due to injected policies. The frequencies of considered cases for each model (with percentages in parenthesis) are given in Table C.2.

C.3 Fine-tuning Parameter Settings

We employed QLoRA [209] adapter on top of attention weight matrices and fine-tuned only the LoRA [177] adapters for 3 epochs. The batch size was set to 2 with an initial learning rate of 2e-5 using a cosine scheduler and the Adam optimizer. We used the checkpoint with the lowest validation loss. Additionally, we set warm up to 0.02, gradient accumulation to 4 and evaluated on validation set 10 times during fine-tuning. We set the rank of matrices for LoRA adapters to 16, the scaling factor (lora_alpha) to 16 and the dropout rate for the adapters to 0.05. Besides, 16-bit mixed precision, bfloat16, was employed for memory efficiency and faster fine-tuning.

CURRICULUM VITAE

Recep Fırat Çekinel

EDUCATION

Degree	Department	Date
Ph.D.	Computer Engineering, Middle East Technical University	2020 - 2025
M.S.	Computer Engineering, Middle East Technical University	2017 - 2020
B.S.	Computer Engineering, Middle East Technical University	2013 - 2017

PROFESSIONAL EXPERIENCE

Title	Institute	Date
Research Assistant	Middle East Technical University	2018 - 2025
Predoctoral Visiting Researcher	University of Tübingen	2023 - 2024
Software Engineer	Aselsan	2017 - 2018

SELECTED PUBLICATIONS

- Cekinel, R. F., Karagoz, P., & Çöltekin, Ç. (2025, January). Multimodal Fact-Checking with Vision Language Models: A Probing Classifier based Solution with Embedding Strategies. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 4622-4633).

- Cekinel, R. F., & Karagoz, P. (2024, December). Explaining Veracity Predictions with Evidence Summarization: A Multi-Task Model Approach. In 2024 IEEE International Conference on Big Data (pp. 6924-6932).

- Cekinel, R. F., Karagoz, P., & Çöltekin, Ç. (2024, May). Cross-Lingual Learning vs. Low-Resource Fine-Tuning: A Case Study with Fact-Checking in Turkish. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 4127-4142).

- Cekinel, R. F., & Karagoz, P. (2022, August). Text-Based Causal Inference on Irony and Sarcasm Detection. In International Conference on Big Data Analytics and Knowledge Discovery (pp. 31-45).