People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
SAAD DAHLAB UNIVERSITY OF BLIDA 1
Faculty of Sciences
Department of Computer Science

Master's dissertation

Speciality : Software Engineering

Presented by :

ECHIKR Abdelghafour

THEME

Health Misinformation Detection on Social Networks

Supervized by :

Dr. Madani Amina USDB

Jury :

Prof. Berremdane Djamila

Prof. Cherfa Imene

Acknowledgements

I want to thank God for everything he has given me, for the resilience to continue on this journey and for all his blessings.

Furthermore, I want to thank Dr. Amina Madani for her advice, expertise, understanding and patience that have had very positive impact on me and made me more comfortable expressing my curiosity about this field and continuing to work on this project.

I am also grateful to my family for all the support they have shown me along my journey up to this point and for giving me a reason to strive and work as hard as I can for the sake of knowledge.

ملخص

أصبح انتشار المعلومات الصحية المضللة على منصات التواصل الاجتماعي مصدر قلق بالغ للصحة العامة، إذ يُقوّض الثقة بالنصائح الطبية، ويُعزز السلوكيات الضارة، ويُضعف الاستجابة للأزمات العالمية مثل جائحة Covid-19. وعلى عكس المعلومات المضللة الشائعة، غالبا ما تستغل الأكاذيب المتعلقة بالصحة اللغة العلمية، والجاذبية العاطفية، والمعرفة الطبية المتطورة، مما يُصعّب اكتشافها وتنظيمها. وتُضخّم منصات التواصل الاجتماعي، بما تتمتع به من سرعة في مشاركة المحتوى وشبكات الصدى، الروايات المضللة قبل أن يتمكن مدققو الحقائق من الرد.

تواجه أنظمة الإشراف اليدوي وأنظمة الكشف التقليدية صعوبة في مواكبة حجم المحتوى وتعقيده، خاصةً عندما تكون المعلومات المضللة خفية، أو مرتبطة بالسياق، أو تُشارك بنية مضللة. وتُشكّل الطبيعة الديناميكية وغير الرسمية للغة وسائل التواصل الاجتماعي تحديات إضافية للنهج القائمة على القواعد أو الكلمات المفتاحية.

ولمعالجة هذه المشكلات، تستكشف هذه الأطروحة حلاً قائما على التعلم العميق باستخدام بنية ModernBERT. ومن خلال ضبط النموذج على بيانات المعلومات الصحية المضللة المصنفة من منصات التواصل الاجتماعي، فإننا نظهر قدرته على التقاط الأنماط السياقية واللغوية التي تميز الادعاءات الكاذبة عن الادعاءات الواقعية، مما يوفر الأساس لأدوات الكشف الآلية القابلة للتطوير بشكل أكبر.

الكلمات المفتاحية: معلومات صحية مضللة، وسائل التواصل الاجتماعي، BERT، ModernBERT، التعلم العميق، معالجة اللغة الطبيعية، اكتشاف المعلومات المضللة، الصحة العامة، نماذج المحولات، التضمينات السياقية.

Abstract

The spread of health misinformation on social networks has become a critical public health concern, undermining trust in medical advice, promoting harmful behaviors, and weakening responses to global crises like the COVID-19 pandemic. Unlike general misinformation, health-related falsehoods often exploit scientific language, emotional appeal, and evolving medical knowledge, making them harder to detect and regulate. Social platforms, with their rapid content sharing and echo chambers, further amplify misleading narratives before fact-checkers can respond.

Manual moderation and traditional detection systems struggle to keep pace with the volume and complexity of content, especially when misinformation is subtle, context-dependent, or shared with misleading intent. The dynamic and informal nature of social media language presents additional challenges for rule-based or keyword-driven approaches.

To address these issues, this thesis explores a deep learning–based solution using a ModernBERT architecture. By fine-tuning the model on labeled health misinformation data from social platforms, we demonstrate its ability to capture contextual and linguistic patterns that distinguish false claims from factual ones, providing a foundation for more scalable, automated detection tools.

Keywords: health misinformation, social media, BERT, ModernBERT, deep learning, natural language processing, misinformation detection, public health, transformer models, contextual embeddings.

Résumé

La diffusion de fausses informations sanitaires sur les réseaux sociaux est devenue un problème majeur de santé publique. Elle sape la confiance dans les conseils médicaux, encourage des comportements néfastes et affaiblit les réponses aux crises mondiales comme la pandémie de COVID-19. Contrairement à la désinformation générale, les fausses informations sanitaires exploitent souvent le langage scientifique, l'attrait émotionnel et l'évolution des connaissances médicales, ce qui les rend plus difficiles à détecter et à réguler. Les plateformes sociales, avec leur partage rapide de contenu et leurs chambres d'écho, amplifient encore les récits trompeurs avant que les vérificateurs de faits ne puissent réagir. La modération manuelle et les systèmes de détection traditionnels peinent à suivre le volume et la complexité du contenu, en particulier lorsque la désinformation est subtile, contextuelle ou partagée avec une intention trompeuse. La nature dynamique et informelle du langage des médias sociaux pose des défis supplémentaires aux approches basées sur des règles ou des mots-clés.

Pour répondre à ces problématiques, cette thèse explore une solution basée sur l'apprentissage profond utilisant une architecture ModernBERT. En affinant le modèle sur les données de désinformation sanitaire étiquetées provenant des plateformes sociales, nous démontrons sa capacité à capturer des modèles contextuels et linguistiques qui distinguent les fausses déclarations des déclarations factuelles, fournissant ainsi une base pour des outils de détection automatisés plus évolutifs.

Mots Clés : misinformation sur la santé, médias sociaux, BERT, ModernBERT, apprentissage profond, traitement du langage naturel, détection de misinformation, santé publique, modèles de transformateurs, vecteurs contextuels.

Contents

List of Figures

List of Tables

General Introduction

Problem statement

The rapid spread of health misinformation on digital platforms, particularly during events like the Covid-19 pandemic, has become a significant threat to public health. With the rise of social media, false health claims can spread faster and reach wider audiences than accurate information, creating what the World Health Organization (WHO) calls an "infodemic" [3], [4].

Misinformation can lead to harmful behaviors, such as vaccine hesitancy or the adoption of unproven treatments, and even incite violence or hate speech against specific groups [5].

The aim of this work is to propose a classification method to detect misinformation in the content of messages related to the health field, exchanged on social networks.

Research methodology

Our research serves to improve the task of misinformation detection on social networks, focusing primarily on text which explains our choice to use the CoAID dataset and to utilize deep learning methods namely ModernBERT which has not yet been trained on health data. That was the reason we have chosen this model to fine-tune it on the common CoAID dataset to solve health misinformation detection tasks.

Dissertation organization

The structure of this thesis will be as follows:

Chapter 01: We will tackle the basics of health misinformation and common approaches to detect it using Artificial Intelligence and Deep Learning techniques.

Chapter 02: The second chapter will contain the state of the art presenting the related work in the field of health misinformation detection, reviewing a few relevant

research papers.

Chapter 03: In the third chapter, we present our contribution and explain how it implements deep learning building on prior work to solve the problem at hand.

Chapter 04: In the fourth chapter, we present the experimentation and the evaluation of our new model alongside the different tools, frameworks and dataset that we used.

Chapter 1

Misinformation detection and AI: an overview

1 Introduction:

Artificial intelligence (AI) is dominating today's headlines, revolutionizing industries with its rapid advancements. Breakthroughs emerge daily, from generative models like ChatGPT [1] to cutting-edge AI in health and autonomous systems. It remains at the forefront of global discourse for reshaping our lifestyle and how we carry out our daily tasks.

AI offers powerful assets when it comes to misinformation detection, tools to analyze textual content, user behavior and propagation patterns. By combining natural language processing (NLP) [6], [7] with deep learning, subtle linguistic signs of deception can be detected, while network analysis aids tracking misinformation. This study leverages these tools to build strong misinformation classifiers, meeting a key digital public need. This chapter explores the application of AI in misinformation detection with a focus on leveraging its models to navigate the challenges posed by entities that aim to spread misinformation.

2 Artificial Intelligence (AI):

AI refers to the simulation of human intelligence in machines designed to perform tasks that usually require human cognition, such as learning, reasoning, problem-solving and decision-making. AI systems leverage algorithms and computational power to analyze vast amounts of data, identify patterns, and make predictions or decisions with minimal human intervention [8]. At its core, AI encompasses machine learning (ML) - where sys-

[1] https://chatgpt.com

tems improve through data experience without explicit programming [9] - and its advanced subset, deep learning (DL), which uses multi-layered neural networks to automatically extract complex patterns [10]. These technologies power modern applications from virtual assistants to medical diagnostics. The exponential growth of AI has three key drivers: explosive data growth, enhanced computing power and algorithmic breakthroughs. Transformative architectures like convolutional neural networks and transformers now achieve human-level performance in vision and language tasks [11].

3 Machine Learning (ML):

Machine learning is a term first introduced by Arthur Samuel in 1959 [12], Samuel described machine learning as the field that enables computers to learn from experience without being explicitly programmed, demonstrated in his checkers playing program.

Traditional Programming, is when the developer or programmer writes explicit instructions (rules) for the computer to execute. The computer is provided with input data and programmed logic to produce an output (See Figure 1.1).

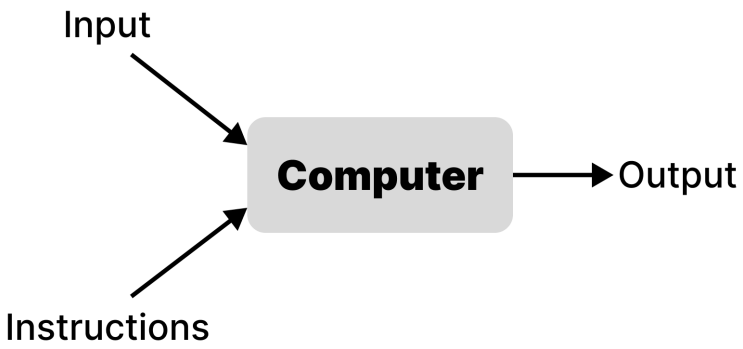

Figure 1.1: Traditional programming principle

Machine Learning is when data input and outputs are provided for the computer so that it can develop its own thinking and reasoning, which is often called 'Model' (See Figure 1.2).

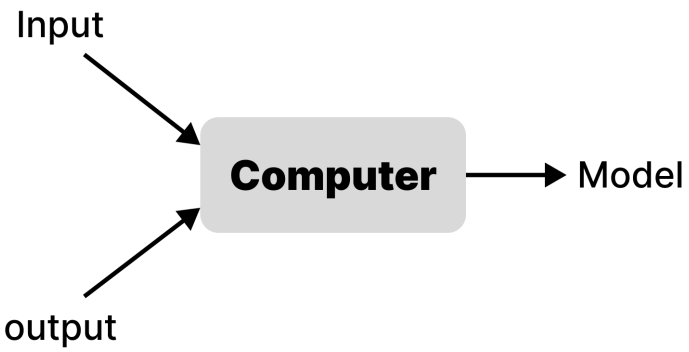

Figure 1.2: Machine learning principle

Machine Learning has four sub-categories, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning and Reinforcement Learning. A brief overview of each one is provided below.

3.1 Supervised learning:

Supervised learning is a type of machine learning approach where a model is trained on a labeled dataset, which means that each training example is assigned a correct output, thus the term 'Supervised'. The main goal of this approach is for the model to learn a mapping from inputs to outputs so that it can accurately predict outcomes for new unseen data. This approach is commonly used in tasks such as classification and regression, where the system learns from past data to make informed decisions. The learning process only stops when a desired level of accuracy is reached on the training data [13]. This method is widely used in spam detection, image recognition, and medical diagnosis. Figure 1.3 represents the supervised learning flow.

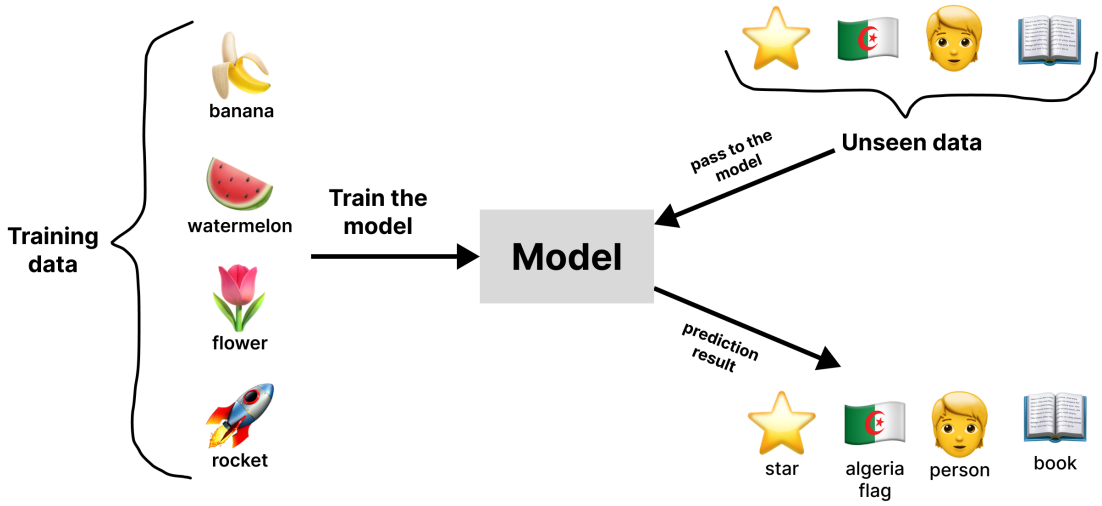

Figure 1.3: Supervised learning flow

3.2 Unsupervised learning:

Unlike supervised learning, unsupervised learning deals with data that has no labeled outputs. It aims to discover hidden patterns within the data without labelled examples. Unsupervised learning algorithms identify structures like clusters by analyzing the input data, common techniques include clustering (e.g., k-means) to group similar data points [14] and dimensionality reduction (e.g., PCA 'Principal Component Analysis') [15] to simplify complex datasets [16]. Figure 1.4 represents the unsupervised learning flow.

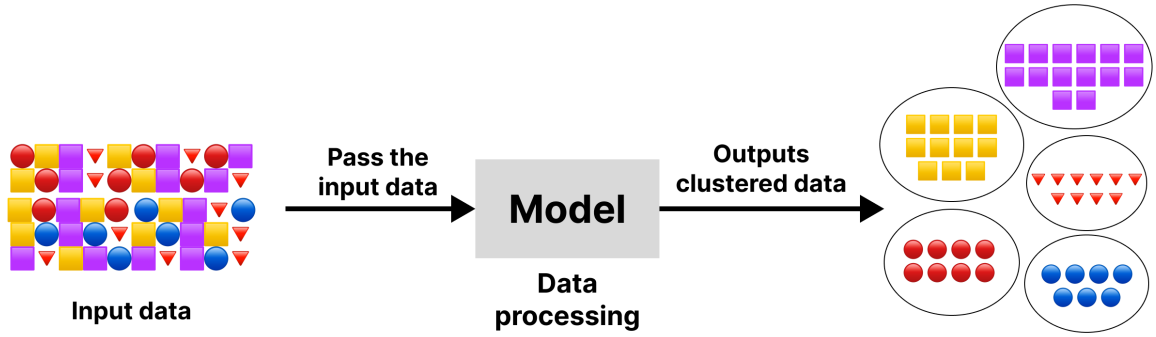

Figure 1.4: Unsupervised learning flow

Unsupervised learning is used in market segmentation, anomaly detection and exploratory data analysis.

3.3 Semi-supervised Learning (SSL):

Semi-supervised learning is somewhat of a middle ground between supervised and supervised-learning, it combines the use of labeled and unlabeled data to train its models more effectively. It uses a small pool of labeled data and a larger pool of unlabeled data, reducing the reliance on expensive labeled data and improving its generalization process making it more efficient, performant and powerful.

3.4 Reinforcement Learning (RL):

Reinforcement learning is a machine learning paradigm where an agent learns decision making strategies through iterative interactions with its environment assisted by a reward signal. RL models follow a principle called "Trial and Error" so that it aims to constantly maximize cumulative rewards. As shown in Figure 1.5, RL models usually consist of:

- Agent: the decision maker
- Environment: the problem space
- Actions: the possible moves that the agent can make
- Rewards: the feedback for the agent's actions

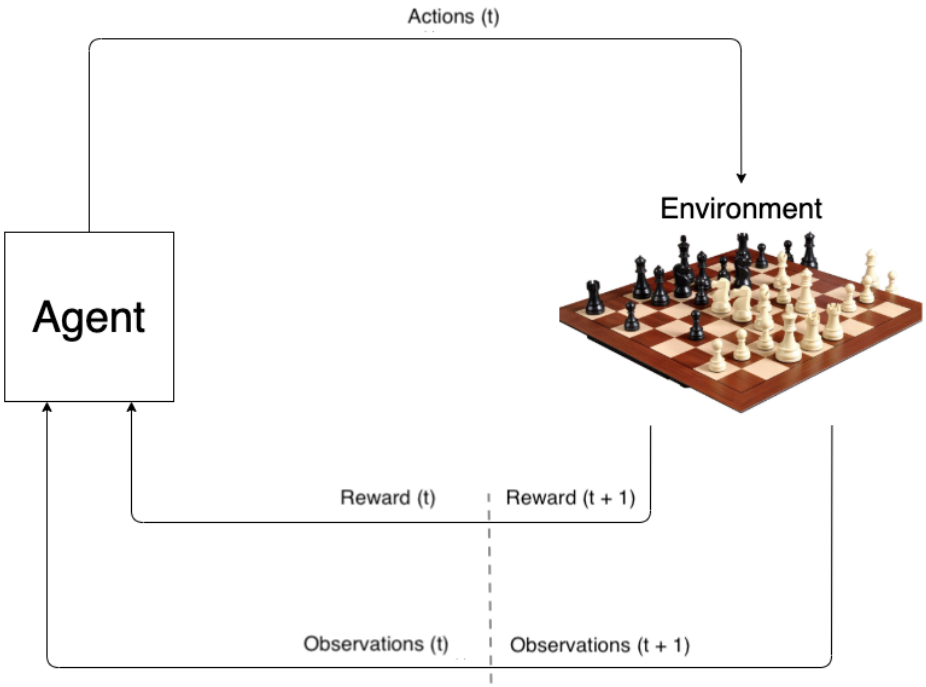

Figure 1.5: Reinforcement learning flow

4 Deep learning:

The term Deep learning began appearing academically in the mid-2000s, through the work of Geoffrey Hinton and his colleagues around 2006 where DL was leveraged in training deep belief systems [17]. Deep Learning is a subset of Machine Learning that uses artificial neural networks with many layers to learn complex patterns from large datasets. It essentially is "computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" [10].

4.1 Convolutional Neural Networks (CNN):

Convolutional Neural Networks, tracing back to Kunihiko Fukushima's neocognitron (1980) [18] are a foundational architecture in DL, particularly well-suited for processing data with spatial hierarchies such as images. They were first implemented by LeCun et al. (1998) [19]. CNNs revolutionized computer vision by automating feature extraction through convolutional filters, leading to breakthroughs in tasks like image classification, object detection, and facial recognition. CNNs have been adapted for applications in medical imaging, video analysis, natural language processing (e.g., sentence classification).

4.2 Recurrent Neural Networks (RNN):

Recurrent Neural Networks date back to John Hopfield's 1982 paper [20], later Michael Jordan (1986) [21] where he described recurrent nets for sequence encoding, and also to Elman, J. L. (1990) [22] where the "simple recurrent network" for temporal sequence learning was introduced. Recurrent Neural Networks process sequential data by maintaining internal memory states that capture temporal dependencies (See figure 1.6). Their recurrent connections allow information to persist, making them well-suited for speech recognition [23], machine translation [24], and time-series prediction tasks. Variants such as Long Short-Term Memory (LSTM) [25] and Gated Recurrent Units (GRUs) [26] overcome vanishing gradients, facilitating the processing of extended sequences.

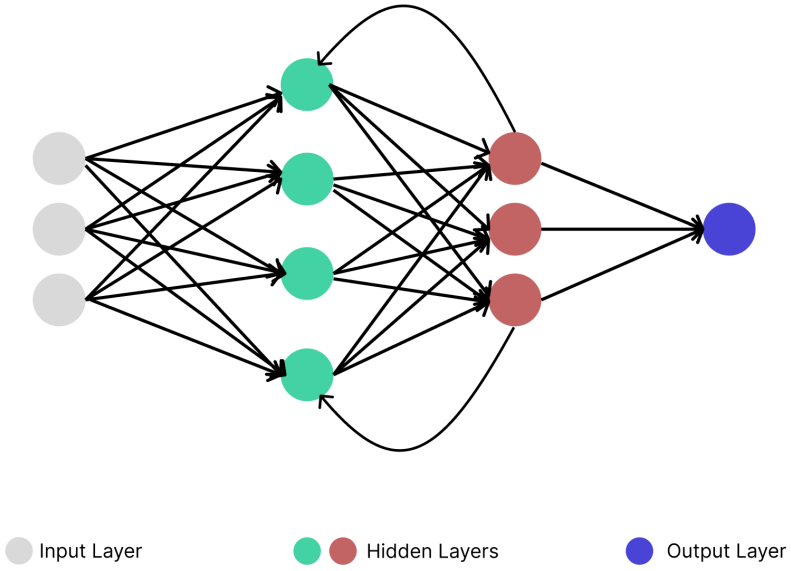

Figure 1.6: Recurrent neural networks structure

4.3 Long Short-Term Memory (LSTM):

Long Short-Term Memory (LSTM) networks were introduced by Hochreiter and Schmidthuber (1997) [25], they represent a specialized type of recurrent neural network (RNN) designed to address the vanishing gradient problem in traditional RNNs. LSTMs can leverage gating mechanisms (input, output, and foreign gates) to selectively keep or discard information over long sequences which makes them highly effective for modeling temporal dependencies.

Key applications of LSTMs include speech recognition [23], where LSTMs outperform conventional HMMs, machine translation through sequence-to-sequence architectures, and time-series prediction in domains like finance and healthcare.

LSTMs have also been pivotal in handwriting recognition [20], it stays widely used despite competition from transformers.

4.4 Transformers:

Transformers, introduced by Vaswani et al. (2017) [1], revolutionized natural language processing (NLP) through their self-attention mechanism, enabling parallel processing of sequential data and capturing long-range dependencies. Transformers process entire

sequences simultaneously unlike recurrent architectures thus making them highly efficient for tasks like machine translation, text generation, and image recognition. Their model architecture is presented in figure 1.7.

Transformer architecture can be seen in models like BERT for bidirectional language modeling and ChatGPT for conversational AI.

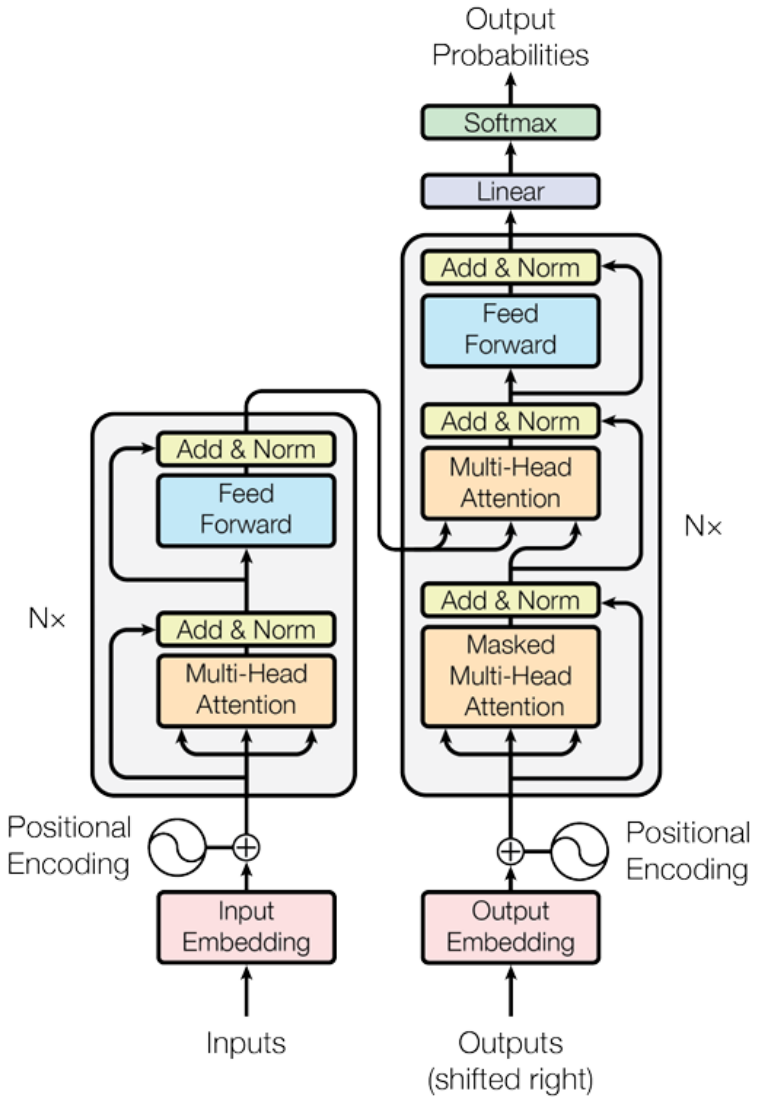

Figure 1.7: Transformer architecture [1]

4.5 Bidirectional Encoder Representation from Transformers (BERT):

Bidirectional Encoder Representations from Transformers (BERT), was first introduced by Devlin et al. (2019) [2], is a transformer-based language model that implements a bidi-

rectional pre-training approach. BERT has achieved state-of-the-art results on numerous NLP tasks by jointly conditioning on both left and right context in all layers. BERT's success has spawned domain specific variants such as BioBERT [27] and LegalBERT [28], which demonstrates its versatility across various fields.

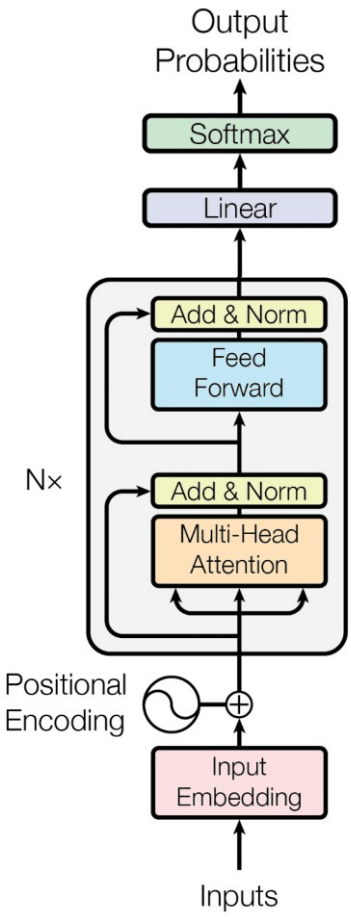

Figure 1.8: BERT architecture [2]

5 Natural Language Processing (NLP):

NLP is a field that enables computers to understand, interpret, and generate human language. What started as rule-based systems has gradually evolved into sophisticated neural architectures that can understand nuance and context behind text. The breakthrough came with deep learning's transformer architecture introducing the attention mechanism which changed the way machines process sequential data, we can effectively

see this applied in machine translation such as Google Translate[2] which is based on the transformer model, BERT model that processes text bidirectionally, understanding context from both directions simultaneously proving to be highly effective for tasks such as sentiment analysis (e.g., classifying product reviews [29]) and chatbots like GPT-3 [30] that's primarily focused on text generation by predicting next words based on preceding context.

These models now demonstrate AI's evolution from simple and narrow task-specific tools, to general-purpose language understanding systems that powers everything from search engines to chatbots.

6 Misinformation detection:

Misinformation detection has become a growing field of research due to the rapid spread of false and misleading content online. Traditional methods rely on manual fact checking, in which trained experts verify claims against credible sources [31]. Modern practical approaches look at how language is used, such as detecting exaggerated claims or conflicting stories or narratives [32]. Some social media platforms now rely on users to report misleading content, but these systems often face problems with bias and manipulation [33]. One of the main difficulties is maintaining a balance between accurately detecting misinformation and protecting freedom of speech, especially in politically sensitive situations.

7 Misinformation detection and AI:

Misinformation, defined as false or misleading claims, poses significant risks to the public, especially on social platforms. Detecting such misinformation requires advanced techniques from Artificial Intelligence (AI), including Machine Learning (ML) and Deep Learning (DL). ML models, such as SVM [34], [35] and Random Forests [36], rely on hand-crafted features, while DL models, such as BERT [2], leverage pre-trained transformers for text classification.

Artificial intelligence and machine learning (ML) are increasingly pivotal in detecting misinformation by analyzing patterns in language, network spread and user behavior [37].NLP (Natural Language Processing) models can flag suspicious claims such as exaggerated "cure-all" promises by comparing them against trusted medical sources [38]. Where AI tools such as Botometer [39], [40] are used to identify automated accounts and coordinated fake activity, which are commonly used to boost the spread of misinforma-

[2]https://translate.google.com

tion [41]. Machine learning algorithms also assess virality risks by tracking engagement metrics, such as sudden surges in sharing or unusual patterns in comments [42]. However, limitations include biases in training data (e.g., mislabeling minority health topics as misinformation) and deceptive attacks where malicious actors subtly evade detection [43] (e.g., misinformation spreaders can tweak or misspell words intentionally to avoid detection, where words like "vaccine" are replaced by "v@ccine"). Hybrid approaches where AI and human judgment are combined together are emerging as a solution to these liabilities, ensuring thoughtful decision making and scalability [44].

8 Common approaches to misinformation detection:

Misinformation detection can be a delicate and complicated matter, since misinformation can take many forms requiring researchers to adapt and develop new models and techniques to tackle it, most of the common approaches can be categorized into the following:

8.1 Content-based detection:

This approach focuses on analyzing the textual or visual content of the message itself. Linguistic features such as sentiment, stance, and syntax are frequently examined to identify misinformation patterns [45]. Some researchers take this a step further by modeling the narrative structure of claims, distinguishing between conspiracy narratives, fake cures, and alarmist messaging. To assess accuracy, they often use semantic analysis methods such as pretrained language models to compare claims against reliable medical sources. Unusual or manipulative use of medical jargons can also raise red flags in this type of analysis.

8.2 User-based detection:

User-based approaches examine characteristics of the user disseminating the information. The user's information often called "Metadata" such as account age, follower/following count and also posting frequency can help detect suspicious behavior [46]. Behavioral analysis has proven to be an important factor to spotting bots or coordinated misinformation campaigns due to their specific posting patterns (e.g., high-frequency posting, low engagement rates).

8.3 Network-based detection:

This approach evaluates how information spreads across a social network, it considers dissemination patterns such as speed, shape and scale of information which proves to be a decisive factor to distinguish factual content from misinformation. Retweet and reply networks can be used to detect groups of users who share content in a coordinated or natural way [4]. Echo chambers and misinformation hubs are often detected by the use of specific community detection algorithms.

8.4 Knowledge-based detection:

Knowledge-based methods involve verifying claims by checking them against trusted sources like fact-checking databases (e.g., Snopes [47], PolitiFact [48]) or knowledge graphs [49]. Named entity recognition (NER) [50], [51] and entity linking techniques are commonly used to match content with factual knowledge sources.

8.5 Multi-modal detection:

Multi-modal detection is when misinformation doesn't only revolve around text but also images or videos. Multi-modal detection incorporates analysis of both text and associated media. For example, re-used or manipulated images from unrelated events can mislead audiences, especially in vaccine misinformation [52].

8.6 Psychological and Sociolinguistic detection:

Emotional and persuasive elements are often seriously taken in consideration by some detection strategies as they are embedded in misinformation. These strategies assess the presence of emotional triggers (e.g., fear, hope, anger, frustration). Misinformation content often exploits cognitive biases, which makes these features valuable for distinguishing between deceptive and credible narratives [53].

9 Conclusion:

In conclusion, AI has a very strong influence on today's world, we have seen how machine learning approaches manipulate data making it a powerful asset, and how deep learning has revolutionized this field completely changing the way we use and search for information. In the next chapter we will take a look at a few research papers relative to health misinformation and how they leverage AI to developing their solutions

Chapter 2

Health misinformation detection: State of the Art

1 Introduction:

The fast growth and popularity of social media networks has enabled the exchange of information between users freely without much restrictions. When it comes to healthcare, unfortunately, we see many users browse for health-related information online only to find a variety of results that mostly stem from unreliable sources that might disseminate incomplete, inaccurate or false information [54]. This chapter will explore the concept of health misinformation and examine a number of approaches and tools employed to detect it. We will also review some related work, highlighting the detection strategies and methodologies adopted in prior research.

2 Health misinformation detection:

2.1 What is health misinformation:

Health misinformation refers to false, inaccurate or misleading health-related claims that are disseminated unintentionally or deliberately, often through digital platforms such as social media, blogs, or forums [55]. Unlike evidence-based medical advice, health misinformation lacks scientific validation and can range from exaggerated claims about treatments to outright conspiracy theories, such as those denying vaccine efficacy or promoting unproven "miracle cures." The consequences of such misinformation are profound, influencing individual health decisions, eroding trust in healthcare systems, and exacerbating public health crises, as seen during the COVID-19 pandemic [42], [56].

2.2 Health misinformation vs health disinformation:

Health misinformation and disinformation are critically distinct, while both involve false or misleading content, misinformation is shared without malicious intent mainly due to ignorance or misunderstanding. (e.g., someone sharing an unverified home remedy for COVID19, believing it to be effective). In contrasts health disinformation is intentionally crafted and disseminated to deceive, often for political, financial, or ideological gain. An example includes coordinated campaigns spreading false claims about vaccines to undermine public health efforts [38], [41]. This intentionality makes disinformation particularly challenging to combat [57], as it is often designed to exploit cognitive biases and emotional triggers.

2.3 Sources and dissemination channels of health misinformation:

Health misinformation thrives across a variety of digital platforms, each contributing to its rapid and widespread dissemination. Social media networks like Facebook, Twitter (now X), and Instagram are primary vectors due to their vast user bases and engagement-driven algorithms, which prioritize sensational content over accuracy [55]. These platforms enable misinformation to spread quickly through shares, retweets, and viral trends, often outpacing corrective efforts. Niche online communities, such as anti-vaccine forums or alternative health groups, further amplify misinformation by creating echo chambers where unverified claims are reinforced without inspection [41]. Additionally, unreliable health websites and influencer endorsements lend false credibility to misleading claims, as they mimic legitimate sources while promoting unproven treatments or conspiracy theories [58]. What is generally considered as a primary barrier for effective management efforts is the decentralized nature of digital media, making cohesive detection efforts more difficult due to the variety of content types across different platforms like YouTube, Instagram, and messaging platforms like WhatsApp [55].

3 General steps to detect health misinformation:

While there are numerous methods for detecting misinformation in general, the process of identifying health misinformation on social networks can be broadly summarized in a few steps:

- ### 3.1 Source Evaluation:

  Check the origin of the information, prioritizing authoritative sources like peer-reviewed studies or official health organizations (e.g., WHO(World Health Organization) [59], CDC(Centers for Disease Control and Prevention) [60], NIH(National Institutes of Health) [61]) to assess credibility, and check for bias and conflicts of interest (e.g., a study on vaping funded by a tobacco company), as seen in viral claims citing obscure blogs over clinical trials [62].

- ### 3.2 Cross-Verification:

  Compare the claims with trusted medical databases such as PubMed [63], Cochrane [64] or UpToDate [65] or scientific literature to confirm accuracy [43] as unsupported claims like "garlic cures COVID" lack endorsement from authorities like WHO or NIH.

- ### 3.3 Content Analysis:

  Examine language for red flags such as exaggerated claims "100% effective", emotional manipulation language "Big Pharma is hiding this", or unsupported causal relationships [66].

- ### 3.4 Algorithmic Monitoring:

  Use AI tools (e.g., Botometer) to track viral trends, bot activity, or coordinated misinformation campaigns across platforms, as seen when false "vaccine side effects" trends surge due to bot activity [41].

- ### 3.5 Fact-Checking:

  Employ fact-checking organizations or expert reviews to validate or debunk health claims [67].

- ### 3.6 User Engagement Review:

  Analyze high-engagement metrics (e.g., rapid shares, likes) to identify potentially misleading content amplified by algorithms [55].

- ### 3.7 Digital Literacy Integration:

  Educate users to recognize misinformation tactics, such as fake experts or cherry-picked[1] data [68].

- ### 3.8 Regulatory and Ethical Alignment:

  Ensure detection methods respect free speech and privacy like GDPR(General Data Protection Regulation) [69] law, while mitigating harm as seen when platforms remove harmful COVID claims while allowing space for legitimate discussion [70].

4 Related work:

This section presents selected related work in the field, showcasing a variety of methodologies and datasets used to implement detection approaches.

The research paper by Barve [71] proposes an approach to detect and classify health misinformation using automated fact-checking using NLP, ML and information retrieval. They proposed a Content Similarity Measure (CSM) algorithm that computes Content Similarity Score (CSS) between URLs and fact-checked references which were classified as legitimate or non-legitimate by following a threshold-based approach. The authors tested their model on the following datasets: CoAID [72], ReCOVery [73], FakeHealth(Story and Release) [74], achieving an accuracy of 87.3%, 89.3%, 85.26% and 88.83% on each dataset respectively when following an algorithmic-approach[2], when following a feature-based[3] approach, their proposed CSM model showed an accuracy of 85.93%, 87.97%, 83.92% and 86.8% respectively, demonstrating superior accuracy over traditional methods like Jaccard [75] and Cosine [76] similarity measures.

The paper by Di Sotto and Viviani [77] contributes research that sought to identify effective features and machine learning techniques to detect online health misinformation on both social media and web pages. The researchers used Classical ML algorithms (Random Forests, Naive Bayes [78], Logistic Regression [79] and Gradient Boosting [80]), Deep Learning algorithms such as Convolutional Neural Networks (CNN) [18] and Bidirectional Long Short-Term Memory networks (Bi-LSTM) [11], [25]. They tested their algorithms on three publicly available datasets: CoAID, ReCOVery and FakeHealth and utilized two

[1] Cherry-picked data is a technique where data is selectively chosen to prove a specific claim, while ignoring other data that contradicts the specified claim

[2] The algorithmic approach refers to using rule-based logic to detect misinformation

[3] The feature-based approach uses specific characteristics or features as input for the machine learning model to classify misinformation

evaluation metrics: AUC (Area Under the ROC curve) [81] and F-measure [82], in addition to Stratified 5-fold cross-validation with feature selection (CFS) [83] for validation. CNNs with word embeddings (CNN(WE)) performed best on the CoAID dataset scoring a precision of 97.3% and 95.3% on AUC and F-measure respectively, whereas ML with word embeddings using ReCOVery dataset scored 92.1% on AUC and 84.8% on F-measure, and ML + TF-IDF [84], [85] (Term Frequency–Inverse Document Frequency) using the FakeHealth dataset scored between 69.3% and 71.7% precision on AUC and between 62.7% and 70.6% precision on F-measure.

The paper by Cui [86] leveraged knowledge graphs (KG) [87] to detect healthcare misinformation to alternate from previous approaches that rely on social contexts, they used a hybrid approach combining Graph Neural Networks (GNNs) [88] specifically Relational Graph Convolutional Networks (R-GCN) [89] with attention mechanisms—and Bidirectional GRUs (BiGRU) [26], [90] for text encoding. They compared their model against eight baselines, including knowledge graph-based methods (KG-Miner [49], TransE [91]), classical text-based models (text-CNN), and social-context-aware systems (dEFEND [92], CSI [93]). The evaluation is conducted on two manually curated datasets (Diabetes and Cancer) using standard metrics: Accuracy, Precision, Recall, and F1-score. Their proposed method called DETERRENT outperforms all baselines, achieving an F1-score of 84.7% (Diabetes) and 93.1% (Cancer), with a 4.8–12.8% improvement over the best competitor [86]. Key innovations include using Bayesian Personalized Ranking (BPR) [94] loss to model negative knowledge graph (KG) relations and improving explainability with attention-weighted KG triples (e.g., "Insulin DoesNotHeal Diabetes"). Limitations include depending on domain-specific KGs and not accounting for time-related changes.

The research paper by Wang [37] proposed a multimodal deep learning model to detect antivaccine misinformation on Instagram, moving beyond text-only methods by combining images, captions, and hashtags. Their model used three branches: a fine-tuned VGG19 [95] network for images, a bidirectional GRU for text (captions and OCR[4]-extracted text), and fastText[5] embeddings for hashtags, all enhanced with a novel semantic- and task-level attention mechanism (SeTa), [98] to focus on key features. They compared their approach against single-modal and other multimodal models on a dataset of 31,282 Instagram posts, achieving 97% accuracy and a 97.3% F1-score outperforming baselines significantly. Key innovations included multimodal feature fusion and an ensemble method combining predictions from all branches. However, some limitations they encountered involved struggles with posts requiring external knowledge (e.g., legal references), OCR

[4]Optical Character Recognition, a technology used to extract text from images, the paper implemented the popular Tesseract OCR algorithm [96]

[5]fastText is a word embedding tool that improves handling of misspelled words and handling text [97]

errors on small text, and a lack of non-English data. Future work suggested adding memory cells for context retention and integrating expert knowledge.

In This paper [99], the authors examined how storytelling (narrative style) affects the spread of health misinformation on Twitter, shifting focus from fact-checking to communication style. They used two COVID-19 datasets (ANTiVax [100] and CMU-MisCov19 [101]), manually labeling 3,000 tweets for narratives (e.g., personal vaccine stories) and training models including logistic regression, BERT variants, and GPT-3 [30] to classify the rest. RoBERTa [102] performed best (F1=0.924). Findings showed narratives boost engagement, even for misinformation (e.g., vaccine conspiracies), especially from influential users. LIWC analysis [103] revealed narratives use more emotional, personal language, while misinformation leans on analytic terms. Key limitations include English-only data and Twitter's changing policies. Unlike KG-based methods, this work highlights how storytelling shapes misinformation reach, suggesting counter-messaging should adopt narrative techniques.

5 Discussion:

Below is a comparison table that lists each paper along with the methodologies used including:

- **Datasets:** which is a structured collection of data that could contain social media posts, online articles or medical claims, that is used to train, validate and test the model. We have mentioned where each dataset has been collected from, its language and its topic.
- **Methods:** which are the techniques used to process data and detect misinformation (e.g., Deep Learning (CNN, Bi-LSTM), Machine Learning (Naïve Bayes, SVM), DETERRENT, CSM).
- **Detection Approach:** is the way the researchers decide to tackle the detection phase, specifying what aspects their models focus on when classifying data. A further detailed description of the detection approaches is mentioned in chapter 1 section 8.
- **Performance:** represents the efficiency of the final model at detecting misinformation accurately with minimal latency, performance is often measured in F1-score, AUC and Accuracy.

Paper	Dataset(s)				Method(s) Used	Detection approach	Performance		
	Name	Topic	Source	Language			Algorithmi Approach (Accuracy)		Feature-Based Approach (Accuracy)
Barve et al.[64]	CredHealth	General health claims	Curated from websites (WHO,CDC, Healthline[1], MayoClinic[2], Wikipedia[3])	English	Content Similarity Measure(CSM)	Knowledge-based +Content-based	0.91		0.882
	CoAID	Covid-19 misinformation	News websites + Facebook, Twitter	English			0.873		0.859
	ReCOVery	Covid-19 misinformation	News websites (e.g.,CNN[4], New York Times[5], Breitbart[6], Natural News[7])	English			0.893		0.879
	FakeHealth (Story)	Health news articles	News articles from mediaoutlets (e.g., Reuters Health[8])	English			0.852		0.839
	FakeHealth (Release)	Health press releases	Press releases fromresearch centers, universities	English			0.888		0.868
Di Sotto And Viviani [70]	CoAID	Covid-19 misinformation	News websites + Facebook,Twitter	English	Deep Learning (CNN, Bi-LSTM) Classical ML (Gradient Boosting, Logistic Regression, Naïve Bayes,Random Forests)	Content-Based + Network-Based+ User-based		AUC	F1
	ReCOVery	Covid-19 misinformation	News websites (e.g., CNN, NewYork Times, Breitbart, Natural News)	English			CNN(WE) (CoAID)	0.973	0.953
	FakeHealth	Health news articles	News articles from mediaoutlets + Press releases from research centers	English			ML(WE) (ReCOVery)	0.921	0.848
							ML(TF-IDF) (FakeHealth)	0.717	0.706
Cui et al. [6]	HealthArticles	Diabetes (False cures, Exaggerated risks)	Snopes[9], HoaxyAPI[10], NIH, Mayo Clinic, WebMD[11], KnowLife triples	English	DETERRENT (KG + GNN +Attention)	KnowledgeBased + Content-Based	Accuracy		F1
	HealthArticles	Cancer (Debunked treatments)	Snopes, Hoaxy API,ScienceDaily, Cleveland Clinic[12], KnowLife triples	English			0.920		0.847
							0.965		0.931
Wang et al.[11]	31,282 Collected Instagram Posts	Antivaccine Vs Provaccine or vaccine-irrelevant	Instagram	English	Multimodal DeepLearning Network (VGG19 + fastText + BiGRU)	Multimodal + Content-Based	Accuracy		F1
							0.974		0.973
Ganti et al. [94]	ANTiVax 3k tweets	Covid-19 vaccinemisinformation	Twitter	English	Logistic Regression	Content-Based + Psychological and Sociolinguistic	F1		
					GPT-3		0.704		
					Naive Bayes		0.718		
					SVM		0.734		
					BERT		0.775		
					TwitterRoBERTa		0.886		
					DistilBERT		0.887		
					DeBERTa		0.893		
	CMUMisCov19(1.4m tweets)	General Covid19misinformation	Twitter	English	RoBERTa		0.910		
							0.924		

Tableau 2.1: An Overview of Relevant Work in Health Misinformation Detection

After reviewing the previous research papers, we observed that most studies rely heavily on the use of COVID-19 datasets primarily focused on processing textual data, particularly in English as some datasets filtered out non-English content during pre-processing. There is limited attention given to visual data or multimodal approaches that combine images and text. As a result, research tends to concentrate on a single platform (mainly Twitter) while other social media platforms receive considerably less attention. Some of the works mentioned above used publicly available datasets, such as: CoAID, ReCOVery or FakeHealth, others used manually curated datasets, where they collect data manually.

6 Conclusion:

In this chapter we provided a few relevant papers containing state-of-the-art models and recent contributions to the field of health misinformation detection. The next chapter contains our proposed model, its implementation steps along with the methodology used to evaluate its performance.

Chapter 3

Conception of a new health misinformation detection model

1 Introduction:

After reviewing the related works previously mentioned, and after conducting further research, we will present in this chapter our own approach to health misinformation detection. We will be fine-tuning a BERT model variant, namely the modernized version of BERT called ModernBERT [104] which has not been previously applied on medical data.

2 The architecture of our model:

The four general steps our model follows to are:

- Data pre-processing
  - Data filtering
  - Tokenization
  - Sequence packing
  - Tokenized text embedding
- Model training
  - Forward pass
  - Compute Loss
  - Backpropagation

  - Weight Update (Optimization)
  - Repeat (Epochs)
- Output head
- Results evaluation

Previous research has indeed implemented BERT to detect health misinformation, and others have implemented domain specialized BERT variants (e.g., BioBERT). However, the new ModernBERT model has only been pre-trained on English text and code which motivated us to fine-tune it for health misinformation detection.

In Figure 3.1 we tried to illustrate our model's architecture by putting together the different components it uses during the fine-tuning and prediction process.

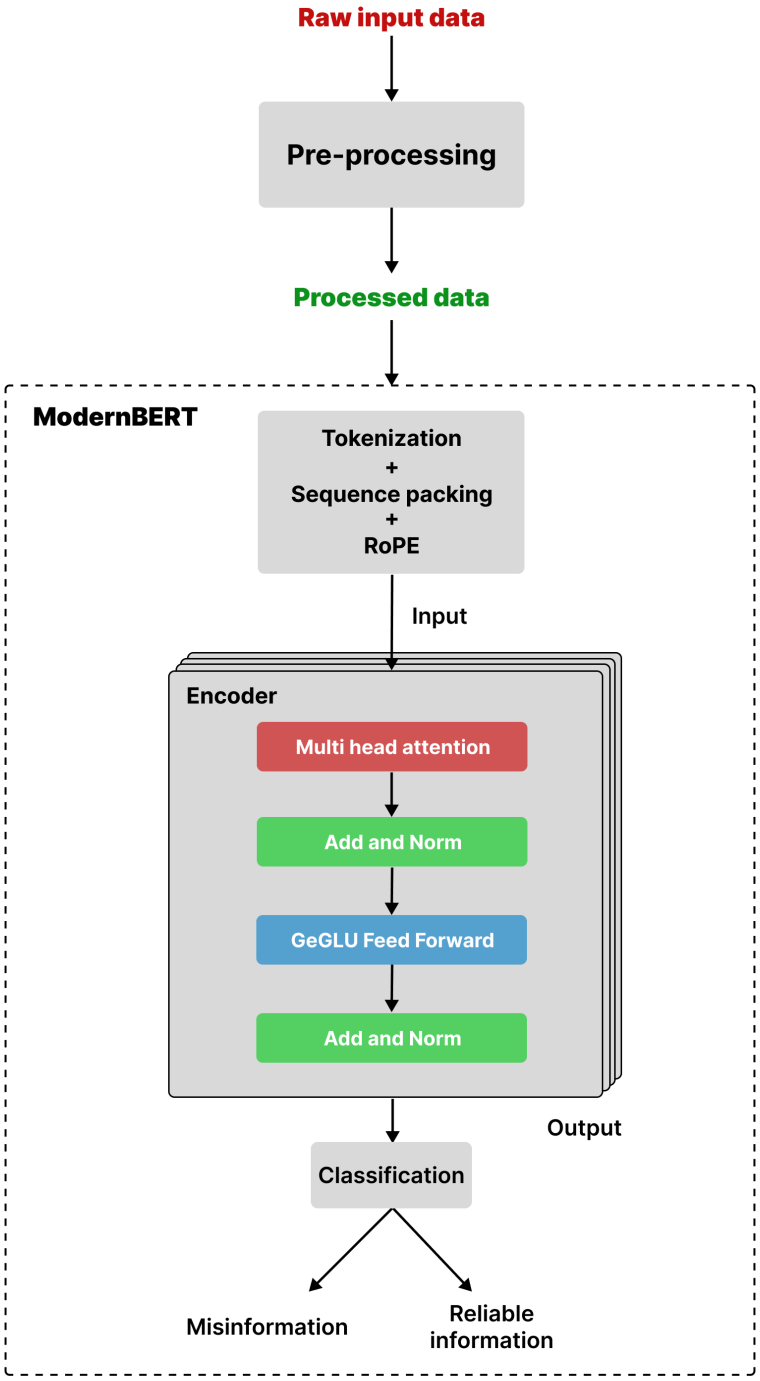

Figure 3.1: Our model's architecture

3 Data pre-processing:

This is considered the primary step to begin the process of fine-tuning our model, as our data must be processed before making other operations. This step is crucial because it

shapes our model's prediction capabilities, and using unfiltered data during training can lead to various issues (e.g., Increased model complexity, inconsistent tokenization, longer training times).

3.1 Data filtering:

The main goal of this step is to simplify, filter and clean the input data before proceeding to the next step which is tokenization. To pre-process our data we have used the following techniques:

- **Cleaning the text:** This is done by filtering out missing values in our data then removing trailing and leading white spaces.
- **Removing irrelevant text:** In order to avoid misinterpretation and confusion, input text must only contain relevant information for our model's training purpose that's why in this step we try to remove:
  - URLs
  - Usernames and hashtags
  - Special characters and emojis
- **Converting text to lowercase:** This step is recommended to reduce vocabulary size and to further optimize tokenization.

A brief example of data cleaning is represented in Figure 3.2 below.

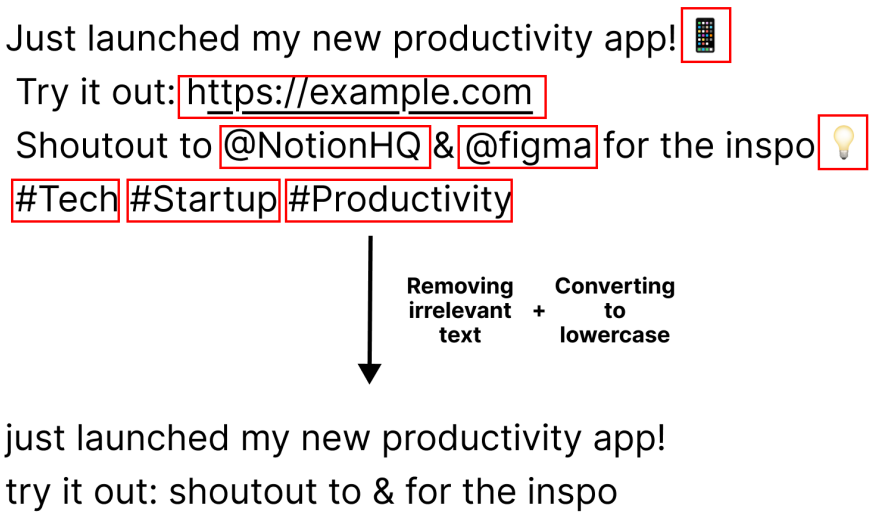

Figure 3.2: Example of data pre-processing

3.2 Tokenization:

After pre-processing our data, the next step is to tokenize it. Tokenization is the process of breaking down sentences into words, sometimes also breaking down complex words, and then assigning IDs to those words before passing it to our model for the training phase.

ModernBERT uses a modern BPE tokenizer, a modified version of OLMo [105] tokenizer providing more token efficiency, while still using the special tokens (e.g., [CLS][1] and [SEP][2]) as the original BERT model. Below is a brief example of tokenization in Figure 3.3.

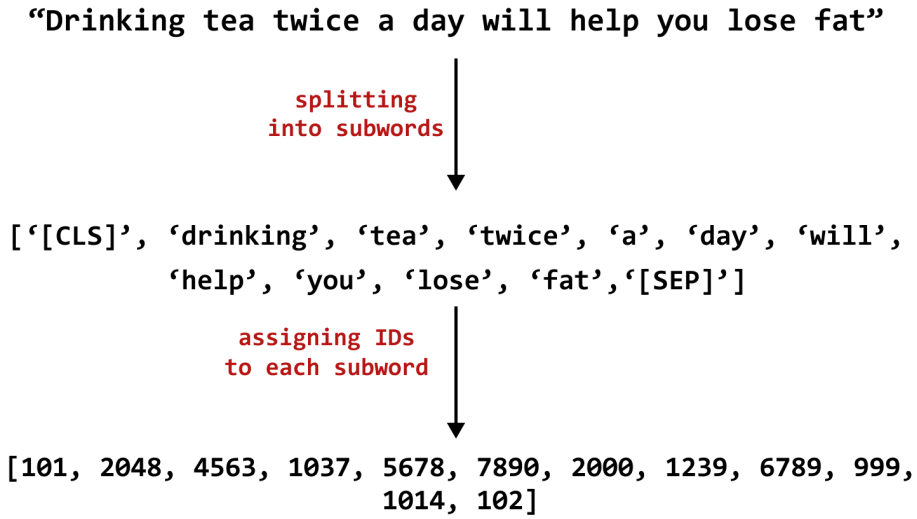

Figure 3.3: Example of tokenization

3.3 Sequence packing:

In addition to tokenization, ModernBERT performs sequence packing, which is the process of concatenating sequences of different length into a single sequence, instead of using padding. This leads to faster training and improved resource economization. An example of this method is represented in Figure 3.4.

[1] [CLS] is a token that signals the start of an input sequence
[2] [SEP] is a token that is used to separate segments in the input

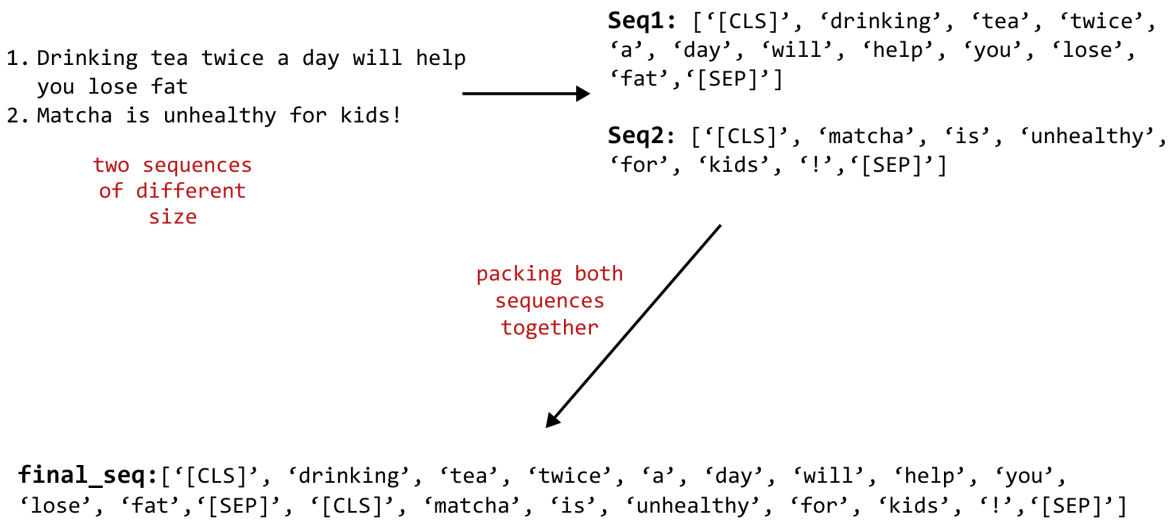

Figure 3.4: Example of sequence packing

3.4 Tokenized text embedding:

Embedding our data is the last step we take before sending it into the model for the training phase, word embedding is the process of mapping each token to embeddings (dense vectors, called Tensors[3]), ModernBERT uses Rotary Positional Embeddings (RoPE) [107] instead of absolute positional embeddings, enabling us to learn context in both directions.

The embedding process consists of the following steps:

1. **Token embedding:** In this step, each token is assigned a unique dense vector (an embedding) of a 768 dimension. Shown in step 1 in Figure 3.5.

2. **Embedding vector rotation:** Each embedding vector is then rotated based on its position in the input sequence using the rotary matrix mentioned above, which makes attention aware of order and distance between tokens. The general rotation formula is as follows:

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \mathbf{R}_{\Theta,m} \cdot \mathbf{W}_{\{q,k\}}\mathbf{x}_m$$

Where the rotary matrix $\mathbf{R}_{\Theta,m} \in \mathbb{R}^{d\times d}$ is block-diagonal[4] with $d/2$ blocks of $2D$

[3] A special type of vector used to stock large sequences of numbers that will take part in complex computations [106]

[4] A block-diagonal matrix has square sub-matrices along the main diagonal and zeros elsewhere.

rotation matrices:

$$\mathbf{R}_{\Theta,m} = \begin{bmatrix} \cos(m\theta_1) & -\sin(m\theta_1) & 0 & 0 & \cdots & 0 & 0 \\ \sin(m\theta_1) & \cos(m\theta_1) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos(m\theta_2) & -\sin(m\theta_2) & \cdots & 0 & 0 \\ 0 & 0 & \sin(m\theta_2) & \cos(m\theta_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos(m\theta_{d/2}) & -\sin(m\theta_{d/2}) \\ 0 & 0 & 0 & 0 & \cdots & \sin(m\theta_{d/2}) & \cos(m\theta_{d/2}) \end{bmatrix}$$

Where:

- $\mathbf{x}_m \in \mathbb{R}^d$: word embedding for the token at position m.
- $\mathbf{q}_m = f_q(\mathbf{x}_m, m)$: positionally encoded query vector.
- $\mathbf{k}_n = f_k(\mathbf{x}_n, n)$: positionally encoded key vector.
- $\theta_i = 10000^{-\frac{2(i-1)}{d}}, \quad i \in \{1, 2, \ldots, \frac{d}{2}\}$ (representing the rotation frequency for the $i-th$ $2d$ subspace of the embedding).

Self-attention is applied here to help keep track of the relationship between tokens, it lets the model know where each token is in the sequence so it can understand token order.

The figure below represents a simple example of the process:

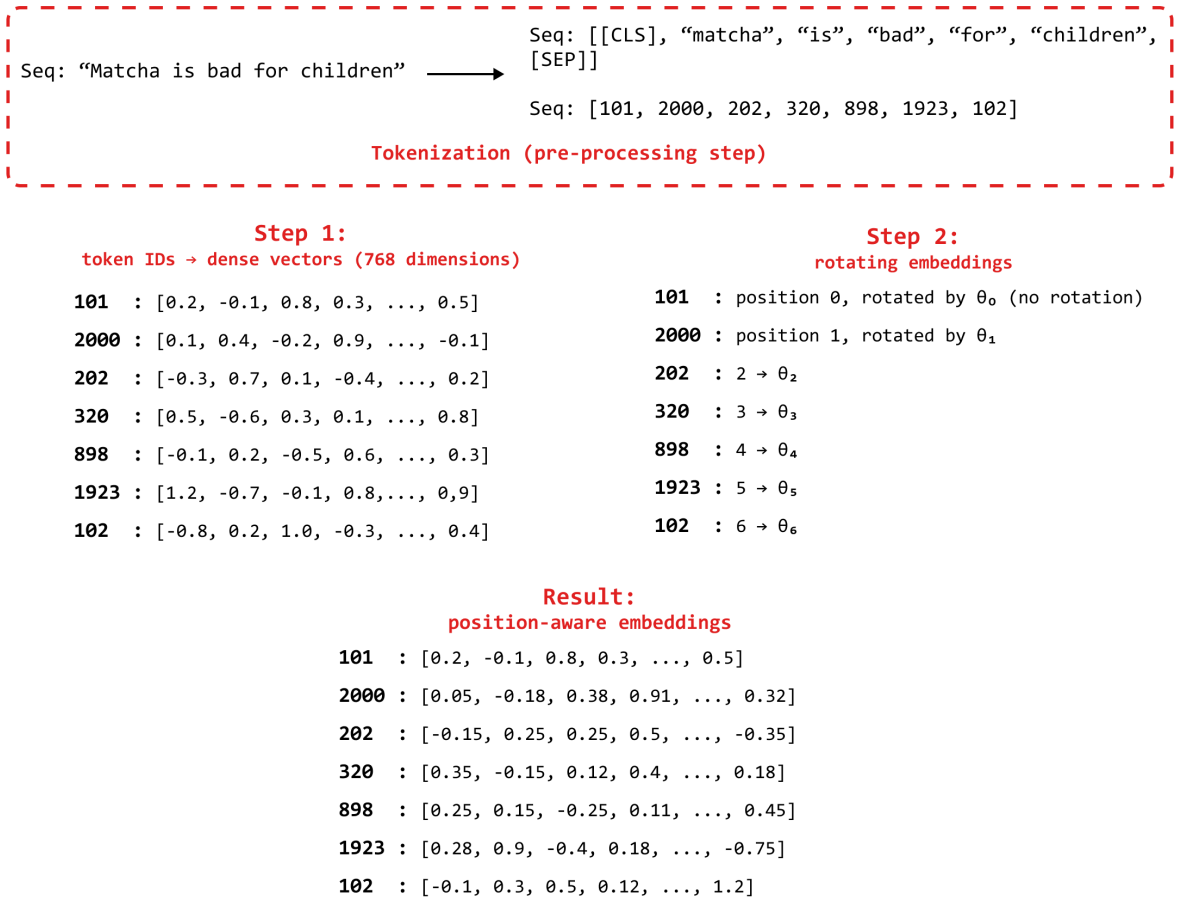

Figure 3.5: Example of the embedding process using RoPE

4 Model training:

The training process happens in the 22 encoder layers of the ModernBERT model, it is the process where our model learns to understand language by adjusting its internal parameters using labeled data.

The first step taken in the training process is splitting our data into validation, testing and training sets (20% testing, 20% validation, 60%training), implementing Stratified sampling [108] to maintain class balance.

This phase only utilizes the training data, leaving aside the testing and validation data for their respective steps.

4.1 Forward pass (Feed forward):

In this step, the tokenized data is passed to the ModernBERT model, which then goes through the model's encoder layers. As shown in Figure 3.1, each layer consists of:

4.1.1 Multi-head attention mechanism:

This is the core self-attention mechanism that allows the model to focus on different parts of the input sequence simultaneously. "Multi-head" means it runs multiple attention operations in parallel.

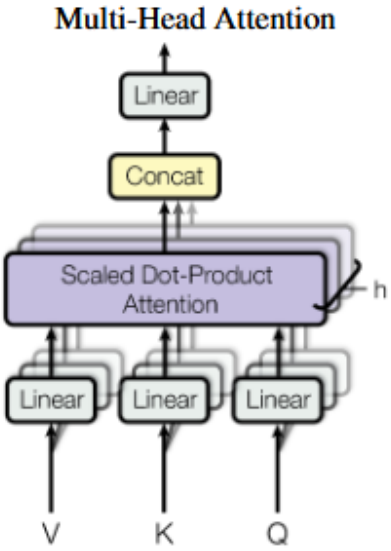

Figure 3.6: Original multi-head attention scheme [1]

ModernBERT builds on the original multi-head attention from the transformer architecture in Vaswani et. al (2017) [1] with a few modifications, the most significant being the alternating attention pattern following recent work on efficient long context models [109], where attention layers alternate between global[5] and local[6] attention.

4.1.2 Add and norm layer (first):

The first Add and Norm layer operates immediately after the multi-head attention mechanism and performs two sequential operations:

- **Add(Residual Connection[7]):** The layer combines the attention output with the original input through element-wise addition. This residual connection maintains the original signal pathway and addresses the vanishing gradient problem commonly encountered in deep architectures, thereby supporting more effective parameter optimization.

[5]Global attention is when in every third layer tokens attends to all other tokens in the sequence.
[6]Local attention is when in other layers (non third layers), tokens attend to other tokens within a sliding window of 128 tokens
[7]Residual connection or "skip connections", is the process of adding the original input directly to the output of a layer.

- **Norm(Layer Normalization):** The summed output undergoes normalization to standardize activation magnitudes across the feature dimensions. ModernBERT employs conventional Layer Normalization [110], which normalizes inputs by computing statistics along the feature axis, promoting training stability through consistent scaling.

This step ensures that representations emerging from the attention mechanism retain both enhanced contextual information and appropriate numerical properties for subsequent feed-forward computation.

4.1.3 GeGLU forward network:

A specific type of feed-forward network that uses the GeGLU (Gated Linear Unit with GELU activation) [111] activation function. The feed-forward network in ModernBERT replaced the traditional ReLU/GELU [112] activation with GeGLU which introduces a gating mechanism inspired by GLUs [113]. Mathematically represented as:

$$\text{GeGLU}(x) = \text{GELU}(xW_1) \otimes \sigma(xW_2)$$

where W_1, W_2 are learnable weights[8], σ is the sigmoid [114], [115] function, and $\otimes$ denotes element-wise multiplication.

4.1.4 Add and norm layer (second):

The second add and norm layer mirrors the first, but is applied after the GeGLU forward network:

- **Add:** GeGLU output combines with the pre-GeGLU input through residual connection. This direct pathway allows training signals to flow efficiently through the network, preventing performance degradation that typically occurs in deeper architectures.
- **Norm:** The combined output is then normalized, ensuring consistent activation scales before the data proceeds to the next encoder layers.

4.2 Compute Loss:

After the Forward pass, a loss is calculated to determine the model's performance so far. The higher the loss value is, the less accurate/performant the Feed forward layer was. As

[8]Weights are numbers the model learns to improve prediction by adjusting them during training based on errors.

shown in Figure 3.7.

4.3 Backpropagation:

Backpropagation represents the process that helps a neural network learn by adjusting its weights. After input data goes through the Feed forward layer and loss is calculated, the model works backwards through the network calculating how much each weight in the network contributed to the error. Then the model uses this information to adjust the weights (parameters) to improve future predictions.

4.4 Weight Update (Optimization):

Weight update is the process where the model adjusts its weights to minimize the loss after backpropagation. It measures how much weights should change using Gradients which are calculated during backpropagation. ModernBERT uses StableAdamW optimizer [116] with a defined learning rate to complete this task.

4.5 Repeat (Epochs):

Epochs represent the number of times we train the model on the entire dataset, where one Epoch means one full pass through the entire dataset, with the model repeating the previous steps and updating its weights each time based on errors from previous passes.

For each epoch, the model processes all batches, calculating both loss and accuracy and tracking training and validation metrics, saving the best results based on validation accuracy.

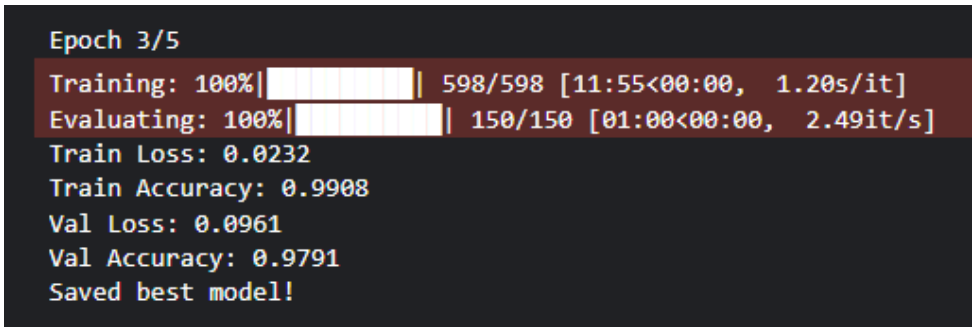

Figure 3.7: Example of training epoch and its results

5 Output head:

5.1 Classification:

The classification process employs a standard approach where the model outputs raw logits[9] for each class through its classification head.

The model then directly applies argmax[10] operation to determine the predicted class where the highest logit value is selected. This proves to be more computationally efficient since the argmax of raw logits is equivalent to the argmax of softmax-normalized probabilities for classification purposes.

6 Evaluation:

Evaluation mainly splits into two parts:

6.1 Training evaluation:

Training is evaluated by measuring our model's efficiency on each epoch during training, this happens on a validation set to see how well our model works on unseen data. This is how we can monitor issues like overfitting[11] and to check if our model is improving after each epoch. The previous figure 3.7 represents this exact step.

6.2 Testing evaluation:

Here we evaluate the performance of the model at prediction when passing it a set of unseen data, which is the testing set previously split from the main processed dataset. The metrics used during this process are detailed in the next chapter.

7 Conclusion:

In conclusion, we have documented the different layers of our model and discussed the role of each one of them in the detection process. In the next chapter we will outline the

[9] Logits are the raw, un-normalized output of a classification model before applying a normalization function like softmax or sigmoid.

[10] https://docs.pytorch.org/docs/main/generated/torch.argmax.html

[11] Overfitting is when the model performs well on training data but poorly on validation data, leaving a gap between Val Accuracy and Train Accuracy values.

results obtained from training our model, alongside the tools and the dataset used in our study.

Chapter 4

Experimentation and results

1 Introduction:

In order to achieve our task at building our misinformation detection model, a set of tools and frameworks is needed. In this chapter we will discuss the dataset we picked and the different tools that went into the creation of our model before discussing the results of evaluating and testing it.

2 Dataset:

The dataset we picked for our research was CoAID[1], which is a diverse COVID-19 healthcare misinformation dataset, including fake news on websites and social platforms, along with users' social engagement to such news.

In total, it contains: 5,216 news, 296,752 related user engagements, 958 social platform posts about COVID-19.

The dataset is broken into four parts, each part represents content collected within a defined span of time, which are:

- **Version 0.1 (05/17/2020):** Representing the initial version of the dataset corresponding to the paper by Limeng et al.(2020).
- **Version 0.2 (08/03/2020):** Represents data collected from May 1, 2020 through July 1, 2020.
- **Version 0.3 (11/03/2020):** Represents data collected from July 1, 2020 through September 1, 2020.

[1]https://github.com/cuilimeng/CoAID

- **Version 0.4 (01/08/2021):** Represents data collected from September 1, 2020 through November 1, 2020

Each of the four folders mentioned above contains three types of files, each one has one of these endings:

- **Tweets.csv:** Containing the tweet IDs.
- **Tweets_replies.csv:** Containing the replies to each of the previous tweets.
- **.csv:** Contains data, its type (post or article), its metadata (e.g., title, newstitle, abstract), its source and its fact_check_url.

Each type of the files above has one of two topics: Claim(Real or Fake), or News(Real or Fake), as shown in the figure below representing one of the dataset's folders and its files.

Figure 4.1: One of the dataset's folders and its structure

We have split our pre-processed dataset into three parts before training, namely:

- **Training data:** Represents 60% of our dataset, used to train the model.
- **Evaluation data:** Represents 20% of the dataset, used to evaluate our model after each training epoch.
- **Testing data:** 20% of the dataset, used to evaluate our model's performance in prediction.

3 Tools and frameworks used:

3.1 Hardware:

Hardware plays a pivotal role in research especially in data intensive fields.

This project was limited by time, but still I have managed to make some progress. I have only used one machine, my personal laptop with an Intel(R) Core(TM) i5-12450H CPU @ 2.50GHz x64 processor with 16 GB RAM.

3.2 Software:

I used the Cursor[2] editor to work on the codebase, it was more than sufficient for this task and to wrap my head around the model's details since I'm already very familiar with it.

However, since the GPU on my machine isn't qualified to train resource intensive models like ModernBERT, I've used Kaggle[3] Notebook for this task. It's very stable and powerful especially the GPU T4x2 that they offer for free was more than enough to take care of the training phase of ModernBERT.

And needless to say, the field standard Python programming language was chosen for this project, due to its simplicity, vast ecosystem and rich libraries and frameworks it made the job much easier and helped me learn many new things along the way.

The implementation also included these frameworks and libraries:

- **Torch (Pytorch)**[4]: Main deep learning framework.
- **Transformers**: (Hugging Face)[5] - For BERT model and tokenizer .
- **Pandas**[6]: For data manipulation and DataFrame operations.
- **Numpy**[7]: For numerical operations.
- **Scikit-learn**[8]: For train_test_split, classification report, Metrics (accuracy, precision, recall, etc.)

[2]https://www.cursor.com
[3]https://www.kaggle.com
[4]https://pytorch.org
[5]https://huggingface.co
[6]https://pandas.pydata.org
[7]https://numpy.org
[8]https://scikit-learn.org

- **Matplotlib**[9]: For creating plots.
- **Seaborn**[10]: For enhanced visualizations (heatmaps)
- **Tqdm**[11]: For progress bars.

4 Evaluation measures:

The model is evaluated at the end of each training epoch as well as after final testing on the held-out test data. A range of evaluation metrics and procedures are employed to assess its performance throughout the training and testing phases.

4.1 Training:

- **Training/Validation Loss:** We use CrossEntropy [117] to calculate the accuracy in both validation and training. The formula is as follows:

$$\mathcal{L}_{\text{train}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i, y_i)$$

- **Training/Validation Accuracy:** Accuracy is calculated as the percentage of the correct predictions, mathematically represented as:

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total samples}}$$

4.2 Testing:

The following measures have been used to evaluate the model's prediction performance:

- **Accuracy score [118]:** Overall correctness.
- **Precision score [119]:** Is the proportion of the predicted positives that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

[9] https://matplotlib.org
[10] https://seaborn.pydata.org
[11] https://tqdm.github.io

- **Recall score [119]:** Is the proportion of actual positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score [120]:** Harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Confusion matrix [121]:** Is a square matrix representing false positives and true negatives, as well as the correct prediction values (true positives and true negatives).
- **ROC Curve [122]:** Is a graph that shows the trade-off between the true positive rate and the false positive rate, reflecting how well the model separates the positive and negative classes.
- **PR Curve [123]:** A plot showing the relationship between precision and recall

5 Model parameters:

5.1 Input shape:

After the data is pre-processed, only relevant information is kept before splitting the dataset into training, validating and testing sets, and passing it to train the model. This is what the input data looks like in a JSON object format, where each attribute represents a column in the final data csv file:


```
{
"text": "u.s. marks deadliest day of coronavirus crisis march 25 2020 more...",
"label": 0,
"date": "05-01-2020",
"source_type": "News",
"has_tweets": 0,
"has_replies": 0,
"file_type": "NewsRealCOVID-19",
"fact_check_url": "webmd.com",
"news_url": "https://www.webmd.com/lung/news/20200325/us-marks-deadliest-..."
}
```

5.2 Training configuration:

We tested our model by making three experiments, passing different parameter values for the trainer, them being:

- **Model name:** We used modernBERT-base for our research.
- **Number of Labels:** This corresponds to the number of classes that data could fit in, in our case there are only two by default as information could be labeled either as reliable (1) or unreliable (0).
- **Batch size:** Is the number of training examples the model sees at once during training before it updates its weights.
- **Learning rate:** The learning rate represents the pace at which the model learns during training. It affects how much the model's weights change after each batch of data is processed, high learning rate can lead to the model learning inappropriately, otherwise low learning rate leads to longer training duration.
- **Number of epochs:** This represents the number of times the model goes through the dataset.

6 Results:

In this section we will discuss the results of experimenting on ModernBERT through fine-tuning and testing on unseen data, providing detailed representation of each run both in numbers and in graphs.

6.1 Fine-tuning:

We have fine-tuned our model three times using the pre-processed CoAID dataset, each time changing the hyperparameters. We stored the training and validation loss, and the training and validation accuracy accross each epoch.

An illustration of the results of the fine-tunning process is provided in Figure 4.2.

Tableau 4.1: Fine-tuning Performance Results Across Different Experiments

Exp	Epoch	Hyperparams	Train Loss	Train Acc	Val Loss	Val Acc
01	1	Epochs:3	0.1664	0.9389	0.1030	0.9582
	2	Batches:8	0.0516	0.9808	0.0925	0.9598
	3	LR:2e-5	0.0209	0.9922	0.0767	0.9724
02	1	Epochs:5 Batches:8 LR:1e-5	0.1493	0.9425	0.1356	0.9540
	2		0.0391	0.9866	0.1114	0.9623
	3		0.0147	0.9958	0.0824	0.9741
	4		0.0128	0.9958	0.0598	0.9799
	5		0.0015	1.0000	0.0822	0.9824
03	1	Epochs:3	0.1343	0.9467	0.0835	0.9665
	2	Batches:16	0.0529	0.9816	0.0607	0.9791
	3	LR:2e-5	0.0172	0.9939	0.1582	0.9540

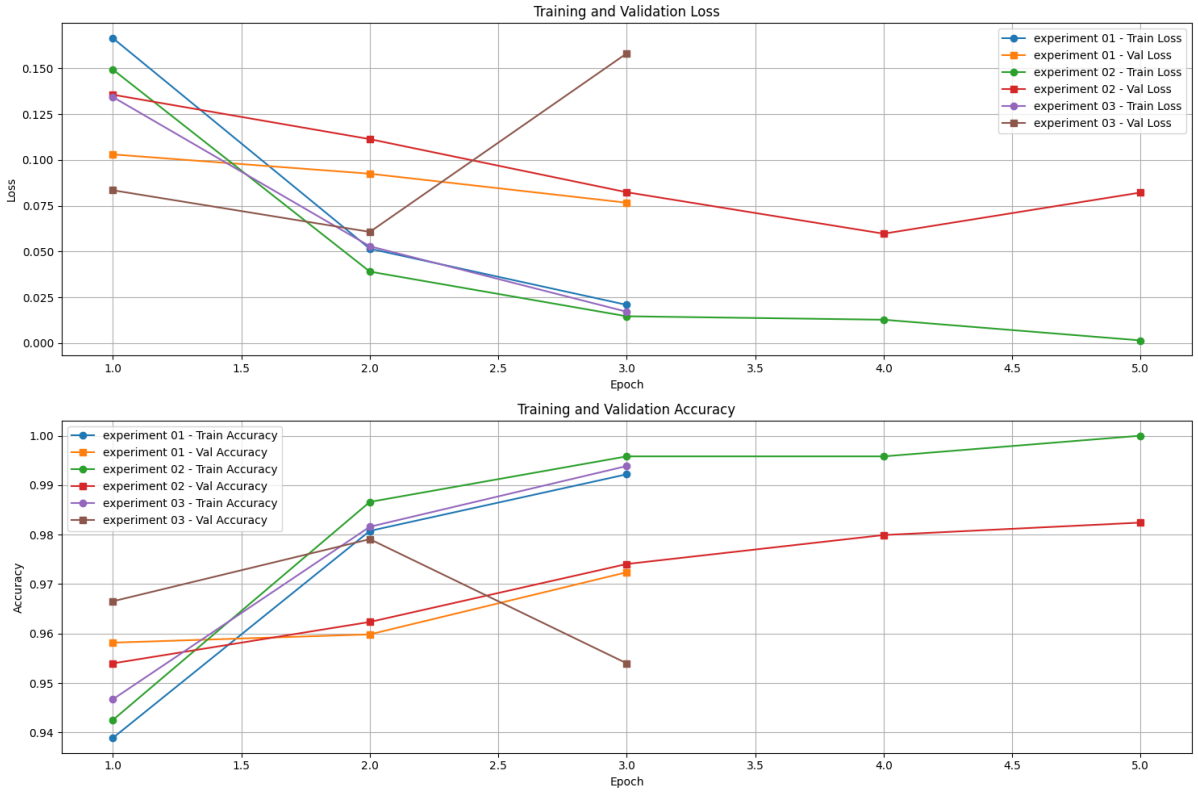

Figure 4.2: Fine-tuning results visualization

6.2 Testing:

Furthermore, we have tested our model on a separate unseen testing data that we split before training, the results are shown below.

Visualized results are shown below, we generated an ROC-curve [115], PR-curve and

Tableau 4.2: Testing Performance Results Across Different Experiments

Experiment	Hyperparameters	Accuracy	Precision	Recall	F1-Score
1	Epochs: 3 Batch: 8 LR: 2e-5	0.9699	0.9330	0.8743	0.9027
2	Epochs: 5 Batch: 8 LR: 1e-5	0.9774	0.9409	0.9162	0.9284
3	Epochs: 3 Batch: 16 LR: 2e-5	0.9732	0.9344	0.8953	0.9144

a Confusion-matrix [121] graphs to illustrate the best model's performance. The best performing model was the one in the second experiment, the ROC-curve in Figure4.3 shows how well it distinguishes between false and reliable information, the PR-curve in Figure4.4 represents the trade-off between decision and recall results.

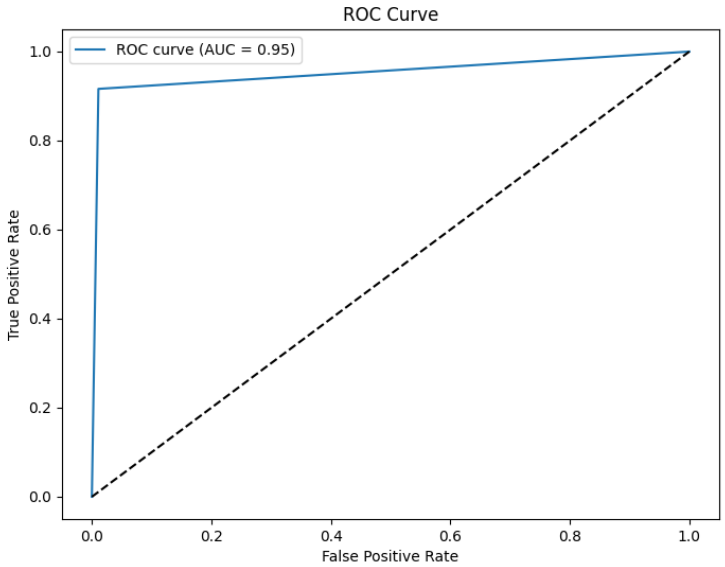

Figure 4.3: ROC Curve - Experiment 2

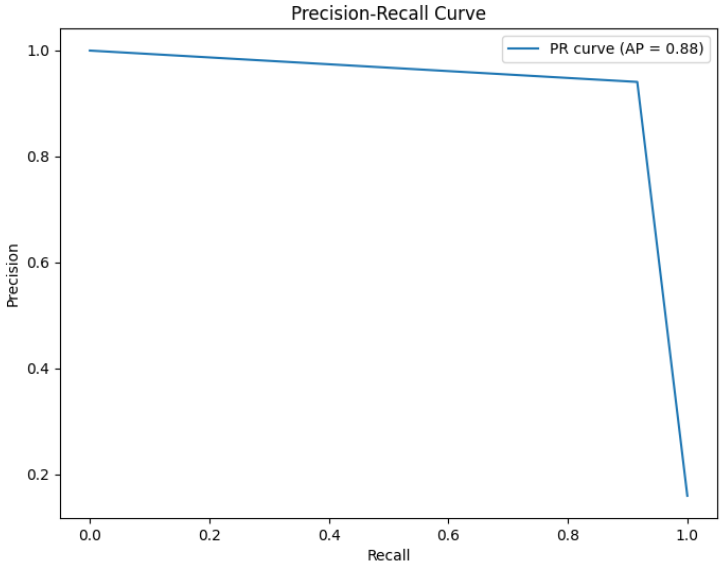

Figure 4.4: PR Curve - Experiment 2

The confusion matrix below represents the amount of false positives and true positives that the model predicted.

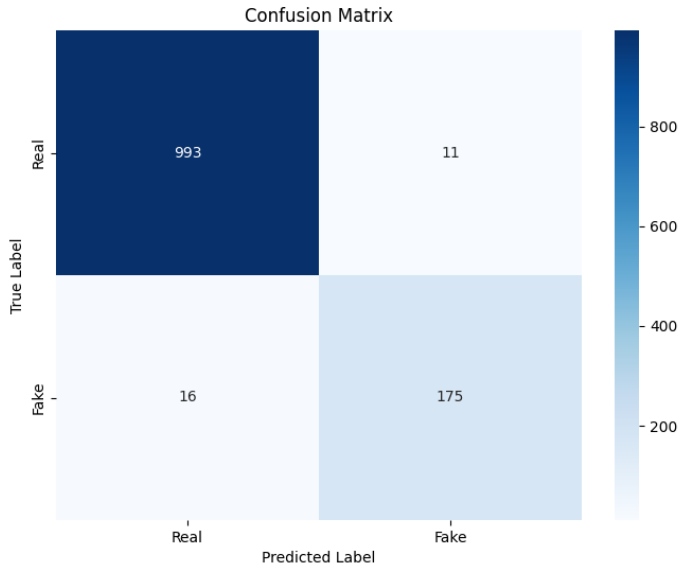

Figure 4.5: Confusion matrix - Experiment 2

We can see that the best model after testing has been the one in the second experiment, using 5 epochs, 8 batch size and $1e-5$ learning rate, achieving the highest performance by all metrics. This can mean that lowering batch size (providing less information to our model per iteration), using slower learning rate and increasing the number of times our

model goes through training data helps our model predict and classify more accurately, proving to be more efficient than other approaches.

7 Comparison:

From the research works mentioned in chapter 2 section 4, two of the five works have used CoAID to test their models, namely Barve et al.(2022) and Di Sotto and Viviani(2022). To facilitate comparison, the table below summarizes the performance of each model alongside our own, evaluated on the common dataset.

Model	**ROC AUC**	**F1-Score**	**Accuracy**
Our Model	0.95	0.92	0.97
Barve et al.(2022)	–	–	0.87
Di Sotto and Viviani(2022)	0.97	0.95	–

Tableau 4.3: Comparison of model performance on the common dataset.

8 Conclusion:

To conclude this chapter, we have evaluated our model's performance in training and in prediction on unseen data using the different evaluation metrics previously mentioned. It was interesting to analyze the model's behavior while experimenting with its hyperparameters. This step proved to be the most challenging. We now proceed to the general conclusion.

General conclusion

In the end, we can say that our research successfully demonstrates the potential of fine-tuning ModernBERT for health misinformation detection using the CoAID dataset. The strong training and validation results achieved indicate that transformer-based models like ModernBERT can effectively learn to distinguish between credible health information and misinformation when properly adapted to domain-specific data.

The fine-tuning process yielded promising performance metrics, suggesting that the model successfully captured the linguistic patterns and semantic features characteristic of health misinformation. These results align with existing literature showing that pre-trained language models, when fine-tuned on specialized datasets, can achieve substantial improvements in domain-specific classification tasks. The CoAID dataset proved to be a valuable resource for this purpose, providing diverse examples of COVID-19 related misinformation that allowed the model to learn robust detection patterns.

However, several limitations constrained the scope of this work. Time constraints prevented the completion of the development of a real-time detection system that could practically implement this approach in live social media monitoring or content moderation scenarios was not feasible within the available timeframe.

These limitations highlight important directions for future work. Developing an efficient real-time implementation would require addressing computational optimization challenges and creating appropriate deployment infrastructure.

Despite these constraints, the positive fine-tuning and testing results provide a solid foundation for future research in automated health misinformation detection. The work demonstrates that fine-tuning modern transformer architectures on specialized health datasets is a viable approach for building effective misinformation detection systems. With additional development time, this approach could potentially be scaled into practical tools for combating the spread of health misinformation across digital platforms, contributing to public health protection efforts in our increasingly connected world.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023. arXiv:1706.03762 [cs].

[2] 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)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[3] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. Mostofa Kamal, S. M. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. Amin Chowdhury, K. S. Anwar, A. A. Chughtai, and H. Seale, "COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis," *The American Journal of Tropical Medicine and Hygiene*, vol. 103, pp. 1621–1629, Oct. 2020.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, Mar. 2018.

[5] I. B. Schlicht, E. Fernandez, B. Chulvi, and P. Rosso, "Automatic detection of health misinformation: a systematic review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, pp. 2009–2021, Mar. 2024.

[6] A. M. Turing, *Computing Machinery and Intelligence*, pp. 23–65. Dordrecht: Springer Netherlands, 2009.

[7] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[8] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Boston: Pearson, third edition ed., 2016.

[9] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, pp. 255–260, July 2015.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[11] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015.

[12] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210–229, July 1959.

[13] E. Alpaydın, *Introduction to machine learning*. Adaptive computation and machine learning, Cambridge, Massachusetts London: The MIT Press, fourth edition ed., 2020.

[14] J. Macqueen, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS,"

[15] K. Pearson, "LIII. *On lines and planes of closest fit to systems of points in space*," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, Nov. 1901.

[16] "Principal Component Analysis for Special Types of Data," in *Principal Component Analysis*, pp. 338–372, New York: Springer-Verlag, 2002. Series Title: Springer Series in Statistics.

[17] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, pp. 1527–1554, July 2006.

[18] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, Apr. 1980.

[19] Y. Lecun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems (NIPS 1989), Denver, CO* (D. Touretzky, ed.), vol. 2, Morgan Kaufmann, 1990.

[20] "(PDF) The A2iA Multi-lingual Text Recognition System at the Second Maurdor Evaluation," in *ResearchGate*.

[21] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*, vol. 121, pp. 471–495, Elsevier, 1997.

[22] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[23] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, BC, Canada), pp. 6645–6649, IEEE, May 2013.

[24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Dec. 2014. arXiv:1409.3215 [cs].

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[26] K. Cho, B. v. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Sept. 2014. arXiv:1406.1078 [cs].

[27] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, Feb. 2020.

[28] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," Oct. 2020. arXiv:2010.02559 [cs].

[29] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, pp. 1–135, July 2008. Publisher: Now Publishers, Inc.

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[31] L. Graves, "Boundaries Not Drawn: Mapping the institutional roots of the global fact-checking movement," *Journalism Studies*, vol. 19, pp. 613–631, Apr. 2018.

[32] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A Stylometric Inquiry into Hyperpartisan and Fake News," Feb. 2017. arXiv:1702.05638 [cs].

[33] G. Pennycook and D. G. Rand, "The Psychology of Fake News," *Trends in Cognitive Sciences*, vol. 25, pp. 388–402, May 2021. Publisher: Elsevier.

[34] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

[35] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[36] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

[37] Z. Wang, Z. Yin, and Y. A. Argyris, "Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2193–2203, June 2021.

[38] B. Guo, Y. Ding, L. Yao, Y. Liang, and Z. Yu, "The Future of False Information Detection on Social Media: New Perspectives and Trends," *ACM Computing Surveys*, vol. 53, pp. 1–36, July 2021.

[39] "Botometer.osome.iu.edu."

[40] K.-C. Yang, E. Ferrara, and F. Menczer, "Botometer 101: social bot practicum for computational social scientists," *Journal of Computational Social Science*, vol. 5, p. 1511–1528, Aug. 2022.

[41] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate," *American Journal of Public Health*, vol. 108, pp. 1378–1384, Oct. 2018. Publisher: American Public Health Association.

[42] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Computing Surveys*, vol. 53, pp. 1–40, Sept. 2021.

[43] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating Fake News: A Survey on Identification and Mitigation Techniques," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–42, May 2019.

[44] A. Kozyreva, S. Lewandowsky, and R. Hertwig, "Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools," *Psychological Science in the Public Interest*, vol. 21, pp. 103–156, Dec. 2020.

[45] 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, pp. 22–36, Sept. 2017.

[46] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, (Hyderabad India), pp. 675–684, ACM, Mar. 2011.

[47] "Snopes.com."

[48] "PolitiFact.com."

[49] B. Shi and T. Weninger, "Discriminative predicate path mining for fact checking in knowledge graphs," *Knowledge-Based Systems*, vol. 104, pp. 123–133, July 2016.

[50] R. Grishman and B. Sundheim, "Message Understanding Conference- 6: A Brief History," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[51] M. Munnangi, "A brief history of named entity recognition," 2024.

[52] F. Alam, S. Cresci, T. Chakraborty, F. Silvestri, D. Dimitrov, G. D. S. Martino, S. Shaar, H. Firooz, and P. Nakov, "A Survey on Multimodal Disinformation Detection," Sept. 2022. arXiv:2103.12541 [cs].

[53] G. Pennycook and D. G. Rand, "Fighting misinformation on social media using crowdsourced judgments of news source quality," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 2521–2526, Feb. 2019.

[54] C. C. Tsai, S. H. Tsai, Q. Zeng-Treitler, and B. A. Liang, "Patient-centered consumer health social network websites: a pilot study of quality of user-generated health information," in *AMIA Annu Symp Proc*, vol. 1137, 2007.

[55] Yvonne Oshevwe Okoro, Oluwatoyin Ayo-Farai, Chinedu Paschal Maduka, Chiamaka Chinaemelum Okongwu, and Olamide Tolulope Sodamade, "A REVIEW OF HEALTH MISINFORMATION ON DIGITAL PLATFORMS: CHALLENGES AND COUNTERMEASURES," *International Journal of Applied Research in Social Sciences*, vol. 6, pp. 23–36, Jan. 2024.

[56] B. G. Southwell, J. Niederdeppe, J. N. Cappella, A. Gaysynsky, D. E. Kelley, A. Oh, E. B. Peterson, and W.-Y. S. Chou, "Misinformation as a Misunderstood Challenge to Public Health," *American Journal of Preventive Medicine*, vol. 57, pp. 282–285, Aug. 2019. Publisher: Elsevier.

[57] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, pp. 1094–1096, Mar. 2018. Publisher: American Association for the Advancement of Science.

[58] A. d. Regt, M. Montecchi, and S. L. Ferguson, "A false image of health: how fake news and pseudo-facts spread in the health and beauty industry," *Journal of Product & Brand Management*, vol. 29, pp. 168–179, Aug. 2019. Publisher: Emerald Publishing Limited.

[59] "World Health Organization (WHO)."

[60] CDC, "Centers for Disease Control and Prevention," May 2025.

[61] "National Institutes of Health (NIH)."

[62] L. Bode and E. K. Vraga, "See Something, Say Something: Correction of Global Health Misinformation on Social Media," *Health Communication*, vol. 33, pp. 1131–1140, Sept. 2018.

[63] "PubMed.ncbi.nlm.nih.gov."

[64] "Cochrane reviews | Cochrane Library."

[65] "UpToDate: Trusted, evidence-based solutions for modern healthcare."

[66] "Index," in *Misinformation and Mass Audiences*, pp. 299–307, University of Texas Press, Jan. 2018. Section: Misinformation and Mass Audiences.

[67] V. Koulolias, G. M. Jonathan, M. Fernandez, and D. Sotirchos, *Combating Misinformation : An ecosystem in co-creation.* OECD Publishing, 2018.

[68] "Journal of Medical Internet Research - eHealth Literacy: Extending the Digital Divide to the Realm of Health Information."

[69] "General Data Protection Regulation (GDPR) – Legal Text."

[70] S. Burris, A. C. Wagenaar, J. Swanson, J. K. Ibrahim, J. Wood, and M. M. Mello, "Making the Case for Laws That Improve Health: A Framework for Public Health Law Research," *The Milbank Quarterly*, vol. 88, no. 2, pp. 169–210, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0009.2010.00595.x.

[71] Y. Barve and J. R. Saini, "Detecting and classifying online health misinformation with 'Content Similarity Measure (CSM)' algorithm: an automated fact-checking-based approach," *The Journal of Supercomputing*, vol. 79, pp. 9127–9156, May 2023.

[72] L. Cui and D. Lee, "CoAID: COVID-19 Healthcare Misinformation Dataset," 2020. Version Number: 3.

[73] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (Virtual Event Ireland), pp. 3205–3212, ACM, Oct. 2020.

[74] E. Dai, Y. Sun, and S. Wang, "Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 853–862, May 2020.

[75] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et du Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, no. 142, p. 547, 1901. Publisher: Imprimerie Corbaz & Comp.

[76] G. H. Thomson, "A HIERARCHY WITHOUT A GENERAL FACTOR," *British Journal of Psychology, 1904-1920*, vol. 8, pp. 271–281, Sept. 1916.

[77] S. Di Sotto and M. Viviani, "Health Misinformation Detection in the Social Web: An Overview and a Data Science Approach," *International Journal of Environmental Research and Public Health*, vol. 19, p. 2173, Feb. 2022.

[78] M. Bayes and M. Price, *An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.* Royal Society of London, Jan. 1763.

[79] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, pp. 215–232, July 1958.

[80] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. Publisher: Institute of Mathematical Statistics.

[81] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 01–11, Mar. 2015.

[82] C. J. Van Rijsbergen, *The Geometry of Information Retrieval*. Cambridge University Press, 1 ed., Aug. 2004.

[83] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper,"

[84] K. Sparck Jones, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL," *Journal of Documentation*, vol. 28, pp. 11–21, Jan. 1972.

[85] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, pp. 513–523, Jan. 1988.

[86] 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*, (Virtual Event CA USA), pp. 492–502, ACM, Aug. 2020.

[87] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, (Vancouver Canada), pp. 1247–1250, ACM, June 2008.

[88] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, Jan. 2009.

[89] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," in *The Semantic Web* (A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, eds.), vol. 10843, pp. 593–607, Cham: Springer International Publishing, 2018. Series Title: Lecture Notes in Computer Science.

[90] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673–2681, Nov. 1997.

[91] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013.

[92] 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*, (Anchorage AK USA), pp. 395–405, ACM, July 2019.

[93] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, (Singapore Singapore), pp. 797–806, ACM, Nov. 2017.

[94] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian Personalized Ranking from Implicit Feedback," May 2012. arXiv:1205.2618 [cs].

[95] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014. Version Number: 6.

[96] "pytesseract: Python-tesseract is a python wrapper for Google's Tesseract-OCR."

[97] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017.

[98] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous Graph Attention Network," in *The World Wide Web Conference*, (San Francisco CA USA), pp. 2022–2032, ACM, May 2019.

[99] A. Ganti, E. A. H. Hussein, S. Wilson, Z. Ma, and X. Zhao, "Narrative Style and the Spread of Health Misinformation on Twitter," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, (Singapore), pp. 4266–4282, Association for Computational Linguistics, 2023.

[100] K. Hayawi, S. Shahriar, M. Serhani, I. Taleb, and S. Mathew, "ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection," *Public Health*, vol. 203, pp. 23–30, Feb. 2022.

[101] S. A. Memon and K. M. Carley, "Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset," 2020. Version Number: 4.

[102] 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," 2019. Version Number: 1.

[103] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker, "The development and psychometric properties of LIWC-22," *Austin, TX: University of Texas at Austin*, vol. 10, pp. 1–47, 2022.

[104] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," Dec. 2024. arXiv:2412.13663 [cs].

[105] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, *et al.*, "Olmo: Accelerating the science of language models," *arXiv preprint arXiv:2402.00838*, 2024.

[106] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[107] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," 2023.

[108] Y. Shi, J. Wang, P. Ren, T. ValizadehAslani, Y. Zhang, M. Hu, and H. Liang, "Fine-tuning bert for automatic adme semantic labeling in fda drug labeling to enhance product-specific guidance assessment," 2022.

[109] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[110] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[111] N. Shazeer, "Glu variants improve transformer," 2020.

[112] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[113] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, pp. 933–941, PMLR, 2017.

[114] P. F. Verhulst, "Recherches mathématiques sur la loi d'accroissement de la population," *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, vol. 18, pp. 1–42, 1845.

[115] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[116] M. Wortsman, T. Dettmers, L. Zettlemoyer, A. Morcos, A. Farhadi, and L. Schmidt, "Stable and low-precision training for large-scale vision-language models," 2023.

[117] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[118] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[119] K. Allen, M. M. Berry, F. U. Luehrs Jr, and J. W. Perry, "Machine literature searching viii. operational criteria for designing information retrieval systems," *American Documentation (pre-1986)*, vol. 6, no. 2, p. 93, 1955.

[120] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 2nd ed., 1979. Introduced the F-measure (F1-score) as a metric combining precision and recall.

[121] K. Pearson, *On the theory of contingency and its relation to association and normal correlation*, vol. 1. Cambridge University Press, 1904.

[122] W. Peterson, T. Birdsall, and W. Fox, "The theory of signal detectability," *Transactions of the IRE professional group on information theory*, vol. 4, no. 4, pp. 171–212, 1954.

[123] C. Van Rijsbergen, "Information retrieval: theory and practice," in *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, vol. 79, pp. 1–14, 1979.