# Exploring the Power of Dual Deep Learning for Fake News Detection

Hounaida Moalla[1,2,*], Hana Abid[1], Dorsaf Sallami[3], Esma Aïmeur[3] and Bassem Ben Hamed[2]
[1]Department of Technological Information, ISET, Sfax, Tunisia
[2]Department of Mathematics and Business Intelligence, ENET'Com, Sfax, Tunisia
[3]Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada
hounaida.moalla@isetsf.rnu.tn, hanaabid@outlook.com, dorsaf.sallami@umontreal.ca,
aimeur@iro.umontreal.ca, bassem.benhamed@enetcom.usf.tn

**Keywords:** Fake news, survey, deep learning, transformers, generative artificial intelligence, detection, personality traits, big five, SEM

**Received:** April 2, 2024

*The rise of social media has intensified the spread of fake news, a problem further exacerbated by generative artificial intelligence (AI). Hence, the need for improved detection of both human-created and AI-generated fake news using advanced AI models is critical. This paper proposes a survey to assess knowledge and attitudes towards news and AI, combining demographic data, personality traits, and the ability to distinguish between real and AI-generated news. Additionally, we create a new dataset, ERAF-News, containing real, fake, AI-generated true, and AI-generated fake news. To classify different types of news, we developed a dual-stream transformer model, DuSTraMo. This model leverages the capabilities of two parallel transformers to enhance the accuracy of news classification. The survey, involving 83 participants from 9 countries, revealed that respondents struggle to differentiate human-generated from AI-generated news. Notably, BERT outperformed GPT-2 and BART in generating realistic text, and RoBERTa and DistilBERT achieved over 98% accuracy in fake news classification. Dual-GPT models also showed high accuracy. This study underscores the effectiveness of the DuSTraMo model and the ERAF-News dataset in enhancing the detection of both human-created and AI-generated fake news. The findings highlight the increasing dominance of AI in this domain and the pressing need for advanced methods to combat fake news. Additionally, a survey examining users' responses to fake news reveals a concerning inability to accurately identify false information.*

*Povzetek: Članek raziskuje zaznavanje lažnih novic z uporabo dvojnoga globokega učenja. Predstavlja nov model DuSTraMo in zbirko podatkov ERAF-News za boljšo detekcijo človeških in AI-generiranih lažnih novic.*

## 1 Introduction

The prevalence of fake news in the modern media environment has grown significantly. It becomes now more difficult than ever to differentiate correct information from false one owing to the fast growth of social media and online platforms [1]. The effects of propagating disinformation are serious, with the ability to change public perception, have an impact on political and social decisions [2], and cause confusion among individuals [3]. The free dissemination of false news can mislead the target audience or gain notoriety or financial advantage. [4, 5].

In the contemporary landscape, interaction with information has undergone a significant evolution, largely attributed to the proliferation of social media and online platforms. Historically, the predominant sources of information were traditional media outlets such as print newspapers [1], radio, and television, all of which adhered to stringent verification protocols.

However, the advent of platforms such as Facebook and Twitter, recently renamed X[2] [3], has revolutionized the way information is disseminated, enabling individuals worldwide to instantaneously share knowledge. Certainly, social media platforms play a significant role in the rapid dissemination of misinformation, complicating the assessment of online information's credibility [6]. This transformative shift not only reshapes the dynamics of information transmission but also underscores the paramount importance of media literacy in discerning truth amidst the continuous influx of online information [7].

Within the persistent issue of fake news, scholars are earnestly committing their efforts to the development of robust methods aimed at identifying and mitigating misleading information [8, 9, 10].

In the literature, several works have been carried out on Fake news detection. These works have varied both on the strategic and technical side and particularly affected

[1] https://rb.gy/5493gt, Last access: 18 Jully 2023

[2] Throughout the rest of the paper, the terms Twitter and X will be used interchangeably.

[3] https://rb.gy/98bc01, Last access: 02 November 2023

three main areas. Firstly, the use of surveys such as [11, 12, 13] have been based either on the analysis of the personality traits of respondents, or the mechanism of circulation of fake news, but have not discussed the degree of awareness of netizens to distinguish between types of news. Secondly, the automatic detection of fake news using ML methods [14, 15, 16] or transformers [15] used simple architectural models. The only recent paper that used dual-stream was [17] but simply duplicated the BERT model in both streams. Thirdly, processing datasets that have just two classes of news (real or fake) [18, 19] besides few works that added the third class of generated news [20, 21, 22]. The authors of [23] maked a multi-classification of information by treating degrees of falsehood such as True, MostlyTrue, Half-True, Barely-True, False, and Pants-on-Fire. No papers proposed to make a classification of four classes (real, fake, generated real and generated fake) and no dataset exists that contains all four classes.

In this paper, the fake news issue is addressed in-depth through the seamless integration of a survey and an artificial intelligence (AI) method. Firstly, a survey is used to delve deep into the public's attitudes. This survey tends to offer quantitative data that explores the nature and extent of the influence of fake news on the behavior of Internet users. Secondly, the capabilities of AI were leveraged, specifically deep learning based on transformers, for the detection of fake news. The innovation in this approach lies in its objective to differentiate between authentic news, false information, and content generated by AI (GAI).

To the best of our knowledge, this is the first paper that explores the synergy between these two approaches and allows for a more thorough examination of the intricate dynamics surrounding fake news, identification, and societal impact.

Hence, the key contributions of this paper are:

- **Diffusing a survey named "Generation of Fake News":** This questionnaire aims to investigate users' personalities, understanding of AI, awareness of fake news, and their capability to distinguish between fake and real news.
- **Creating a new dataset named "Extended Real And Fake News Dataset (*ERAF-News*):** The dataset includes real news, fake news, and real and fake news generated by AI. It can be used to test and refine state-of-the-art algorithms to detect fake news. The idea behind adding the fourth class is to distinguish between generated Real and generated Fake. This allows to optionally accept generated Real News and reject generated Fake ones.
- **Proposing a new Dual-Stream Transformers Model *(DuSTraMo)*:** This architecture aims to exploit complementary capabilities from two parallel models, potentially leading to improved performance on various natural language processing tasks.

The remaining parts of his paper proceed as follows: Section 2 provides an overview of three key areas: (a) prior survey-based research on fake news; (b) studies employing deep learning techniques based on transformers to classify the news; and (c) dynamic landscape of generative AI tools and their impact on fake news generation. Section 3 offers an overview of transformers. Section 4 intricately delves into the specific contributions and explains the methodology. It includes (a) an in-depth exploration of the survey, covering design rationale, section selections, presentation of collected data, and the chosen evaluation method; (b) a comprehensive account of the creation of the novel *ERAF-News* dataset; and (c) an exposition of the innovative *DuSTraMo* architecture designed for the detection of fake news. Section 5 offers an in-depth exploration of the obtained results, encompassing: (a) the findings from the investigation; (b) a performance evaluation of the *ERAF-News* dataset generation; and (c) a comprehensive discussion of the results derived from the deep learning models developed.

## 2 Related works

This section presents the literature in three distinct areas: (1) research focused on surveys [4] in the fake news context, (2) generation process based on a few pre-trained models, and (3) new method for fake news detection.

### 2.1 Surveys for fake news catch

The fast transmission and persistence of misinformation online pose a multifaceted threat [25, 24], impacting various domains including politics [27, 26], culture [29, 28], finance [31, 30], and psychology [24]. For the purpose of evaluating the accuracy of Internet users' answers, conducting an online survey can prove invaluable for assessing the attitudes, cultural orientations, and more of Internet users, thereby measuring the consistency of their responses.

The authors [11] are interested in the different mechanisms of circulation of disinformation on online social networks, and propose avenues for identification and in-depth analysis.

Linguistic analysis of information plays a principal criterion in the detection of false information. For this, one approach is to use a survey comprising a series of questions, [5], which can range from general queries to personalized or domain specific.

The objective of a survey [32] is to explore the correlation between the optimal user experience while navigating social media platforms, the behavior of sharing fake news, online trust, and heightened social media consumption. This research draws upon pertinent studies from the literature on fake news, online trust, and social media usage to inform the development of the questionnaire. Table

[4] Throughout the rest of the paper, the terms survey and questionnaire will be used interchangeably.

[5] https://shorturl.at/aIKP6, Last access: 02 September 2023

Table 1: Comparison of some fake news surveys

| Ref. | Title | Subject | Nb responders |
|---|---|---|---|
| [33] | The role of social media in spreading panic among primary and secondary school students during the COVID-19 pandemic: An online questionnaire study from the Gaza Strip, Palestine | COVID-19 | 942 |
| [34] | Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking | Fake news | 69 |
| [35] | The Influence of Political Ideology on fake news Belief: The Portuguese Case | Fake news on political ideology | 712 |
| [36] | Fake news: the impact of the internet on population health. | Fake news of health information | 1195 |
| [37] | fake news Reaching Young People on Social Networks: DistrustChallenging Media Literacy | Fake news | 408 |
| [38] | Fake or real news? Understanding gratifications and personality traits of individuals sharing fake news on social media platforms | Fake news | 221 |
| [39] | The Role of Risk Perception and Ability to Detect fake news in Acceptance of COVID-19 Vaccine among Students of Shiraz University. | COVID-19 | 382 |
| [40] | Anger makes Fake news viral online. | Fake news (offline) | 1291 |
| [41] | Examining the role of emotions, sharing motivations, and psychological distance of COVID-19-related fake news | COVID-19 | 150 toilet paper shortage-related |
| | | | 149 celebrity scandal rumors |

1 provides a brief overview of the studies on fake news surveys.

Comprehending the elements that influence how individuals react to fake news is vital for devising efficient strategies to minimize its detrimental impact on society. Personality traits play an essential role to shape people's behaviors[12, 13]. The Big Five model [6] seems a highly valuable standard for comprehending personality. This model is commonly named OCEAN (Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism), derived from its five key dimensions. It assesses an individual's position on each dimension, providing valuable insights into their personality traits and behavioral characteristics.

## 2.2 Methods for detecting fake news

Popular techniques for fake news detection, such as source verification, fact-checking, and cross-checking [42] are important to assess the credibility of news stories [43]. However, deeper analysis based on AI and machine learning (ML) algorithms has made it possible for sentiment analysis and language pattern identification, letting specialists to identify disinformation [44, 45].

Academic studies propose diverse techniques for automatic fake news detection, primarily relying on content-based techniques [46, 47], natural language processing (NLP) [18, 48], or data mining [49], ML methods [14, 15], or the use of social context-based techniques, as discussed in [50]. Additionally, some models adopt a hybrid integration of both approaches, as demonstrated by [51].

For assessment and classification of information, the use of NLP techniques and AI methodologies such as random forests (RF) [52], support vector machines (SVM) [53], long short-term memory (LSTM) neural networks [54], and transformers [19, 17, 55, 56], allows information professionals to enhance their ability to detect fake news effectively.

Researchers in [16] propose an innovative language-agnostic technique based on text attributes to discern fake news using various datasets. The outcomes reveal that the RF and SVM algorithms achieve accuracy of around 88% and 89% respectively, when applied to the FakeBr-

[6] https://shorturl.at/abFS4, Last access: 25 September 2023

Table 2: A comparison between fake news detection approaches

| Date | Ref. | Dataset | Dataset classe | | | Models |
|---|---|---|---|---|---|---|
| | | | Real | Fake | GAI | |
| 2019 | [20] | RealNews(Common Crawl) | x | x | x | GPT2, BERT, Grover-Mega |
| 2020 | [18] | FakeOrRealNews | x | x | | ML models |
| | [19] | FakeOrRealNews | x | x | | BERT, RNN |
| | [16] | TwitterBR, FakeBrCorpus, FakeNewsData1, FakeOrRealNews, btvlifestyle | x | x | | ML models |
| | [21] | newsQA dataset extension | x | x | x | Grover-Mega, Zero-shot |
| 2021 | [15] | Fake or real news, Combined corpus | x | x | | ML models, CNN, LSTM, HAN, BERT, RoBERTa, DistilBERT, Electra |
| | [22] | S, xl, s-k, xl-k, Webtext, GPT3-WebtextrealNews | x | x | x | GPT2, GPT3, Grover |
| 2022 | [42] | Debate , PHEME | x | x | | NLP framework |
| | [43] | Amazon, Common Crawl, Fake and Real News | x | x | | Proposed pipline |
| | [44] | Created dataset | x | x | | ML models |
| | [52] | FakeNewsNet | x | x | | ML models |
| | [62] | RAWFC, LIAR-RAW | x | x | | ML models, CNN, RNN, SentHAN, DeClarEdEFEND, SBERT-FC, GenFE, GenFE-MT, CofCED |
| 2023 | [17] | FakeNewsNet | x | x | | Dual BERT |
| | [56] | Private dataset | x | x | | Keyword-based, Rule-based, ML models |
| | [63] | CovidNews +NQ-1500 | x | x | x | BM25, GPT-3.5, GENREAD, REIT, CTRLGEN, REVISE, RoBERTa-based |

Corpus dataset [16]. Furthermore, [56] presents a machine-learning-centered method. Emphasizing the significance of source reliability, this technique attains a commendable accuracy rate of 90%. Table 2 provides a concise overview of the primary studies conducted in the literature concerning the automated detection of fake news [7].

Deep learning single models exhibited superior accuracy in detecting GAI compared to machine learning overall [57]. The authors of [58] performed the classification of Indonesian fake news using pre-trained models based on a multilingual transformer model (XLM-R and mBERT) combined with a BERTopic model as a topic distribution model. Their model provided an accuracy of 0.9051.

However, these models did not delve into the specifics of the generated news and the potential multi-class outcomes it could produce. Indeed, the previous researches did not establish the distinction between the different AI-generated classes, hence the need to generate a dataset containing the GAI-Fake and GAI-True classes. Effectiveness might arise from the intricate attributes of AI-generated content, subtle writing approaches, and the dynamic landscape of misinformation. Moreover, relying solely on single models [60, 61, 59] is proving insufficient. Addressing these limitations, more advanced and adaptable techniques, such as a dual-stream model, which is proposed in Section 4.3, could enhance accuracy and effectively counter the evolving challenges of fake news. Past research has

[7] the words True and Real will be used intrinsically. Likewise, for the words False and Fake.

predominantly relied on machine learning models or a single transformer model, except for the very recent paper [17], which proposed one dual model technique.

In order to classify and understand the nature of news (Fake, Real, GAI-True or GAI-Fake), one can check whether the news contains false, misleading or deliberately false information by: (1) ensuring the credibility of the news [64] even if it is generated by AI; and (2) using automatic detection tools [65] to distinguish authentic news from fake or AI-generated news.

### 2.3 Influence of generative AI on the proliferation of fake news

Generative AI is a sub-field of AI that focuses on developing models and algorithms capable of generating human-like content, such as text, images, and audio. Recently, text generation, in particular, has seen remarkable advancements, thanks to models like GPT-3 (Generative Pre-trained Transformer 3) [66] and its successor GPT-4 [67].

Generative AI has unfortunately been exploited for the generation and dissemination of deceptive news content. Ironically, advancements in technology aimed at detecting false information have accelerated the creation of increasingly sophisticated and perplexing fake news. The demarcation between reality and fabrication has become increasingly blurred with the advent of AI-powered text generation [68]. Indeed, the potency of generative AI possesses a dual nature: it enables the automatic generation of text closely resembling human writing, while, on the other hand, it significantly contributes to the generation of disinformation. In fact, by inputting a prompt or topic, these models can generate seemingly authentic news articles, reports, or social media posts that are entirely fictional [69, 70][8]. As per findings from researchers [9], the utilization of generative technology may render disinformation more cost-effective and simpler to generate, thereby contributing to an increase in the prevalence of conspiracy theorists and the dissemination of false information.

In a study conducted by researchers at OpenAI [71], a cautionary note was issued regarding the potential for the chatbot service to lower the costs associated with disinformation campaigns. It was suggested that malicious actors might be incentivized by the prospect of financial gain, advancing specific political agendas, or sowing discord and confusion. Furthermore, just two months after its launch, ChatGPT faced criticism from the NewsGuard platform [10], a specialized entity in the detection of fake news. NewsGuard noted that the chatbot had the potential to morph into a "superspreader" of disinformation. In an experiment detailed on their platform [11], the NewsGuard team instructed ChatGPT to compose articles that mirrored the viewpoints of prominent conspiracy theorists or biased news outlets. The results of the study uncovered that, out of the 100 misleading stories pre-identified, ChatGPT generated false news for 80 of them. These misleading narratives could appear compelling and authoritative to uninformed readers, despite their basis in falsehoods.

In response to this challenge, researchers have dedicated significant efforts to develop methods for detecting AI-generated content used for deceptive purposes. These methods encompass a variety of approaches, from employing machine learning models trained to recognize patterns in AI-generated text to linguistic analysis that seeks anomalies in language usage [72, 73, 74], including the architectures of transformers, which are elaborated in the subsequent section.

## 3 Background: transformers

Within this section, a brief overview of the architectural structures that underlie transformers will be provided.

### 3.1 Architecture

The transformer architecture, as elucidated in [75], adopts an encoder-decoder structure as shown in figure 1.

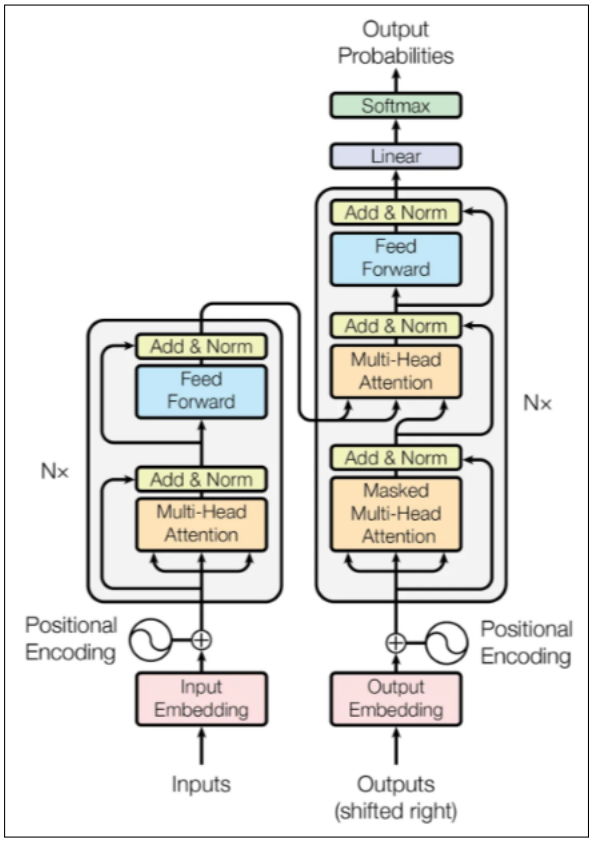



Figure 1: Original structure of transformer

[8] https://shorturl.at/boAHW, Last access: 21 August 2023
[9] https://shorturl.at/nKPS2, Last access: 21 August 2023
[10] https://www.newsguardtech.com/, Last access: 12 September 2023
[11] https://shorturl.at/auwTU, Last access: 12 September 2023

Transformer operates by taking an input sequence $X = (x_1, \ldots, x_N)$ and transforming it into a latent representation $Z = (z_1, \ldots, z_N)$. Notably, due to the auto-regressive nature of this model, the output sequence $Y_M = (y_1, \ldots, y_M)$ is generated one element at a time. In other words, each word $y_M$ is generated using the latent representation $Z$ and the previously created sequence $Y_{M-1} = (y_1, \ldots, y_{M-1})$.

Both the encoder and the decoder components of the transformer employ an identical multi-head attention layer, which plays a pivotal role in data processing. In this mechanism, a single Attention layer maps a query $Q$ and keys $K$ to a weighted sum of the values $V$. For practical reasons, a scaling factor of $\sqrt{\frac{1}{d_k}}$ is introduced to ensure effective operations:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This attention mechanism is fundamental to the transformer's ability to capture complex relationships and dependencies in the input data, enabling its outstanding performance in various natural language processing tasks.

## 3.2 Transformers list

There are numerous transformer models developed for various natural language processing (NLP) tasks. Here are some notable ones:

- **Generative Pre-trained Transformer 2 (GPT-2):** a powerful natural language processing algorithm that can generate text that is both fluent and relevant to the context. It is used in a variety of NLP tasks, such as content creation, chatbots, question answering, sentiment analysis, and text summarization. GPT-2's ability to produce text that is similar to human-written text makes it a valuable tool for a wide range of language-related applications [76].
- **Bidirectional and Auto-Regressive Transformers (BART):** a sequence-generating paradigm that works well for producing text. In order to produce material that flows and makes sense, it can take into account both the context that comes before and after. BART is frequently utilized for machine translation, generative text synthesis, automated summarizing, and paraphrasing [77].
- **Bidirectional Encoder Representations from Transformers (BERT):** a powerful language model that can learn the meaning of words and sentences by reading them from both sides. This makes it very good at understanding the meaning of language and performing many different tasks, like classifying text, recognizing named entities, analyzing sentiment, and understanding context [78].
- **Distillable BERT (DistilBERT):** a smaller and faster BERT. It works just as well as BERT, but it uses less computing power and memory [79].
- **Robustly optimized BERT approach (RoBERTa):** a version of BERT that has been improved to perform better during pre-training. It is often used for the same tasks as BERT, but with better accuracy because of the changes made to its training process [80].

# 4 Proposed approach

In this section, we propose a three-part approach, namely creating an online survey, creating a new dataset, and exploring multiple dual models for classification and detection of fake news.

## 4.1 Online survey

We propose a survey with the comprehensive aim of investigating public perceptions and experiences related to fake news. The objective is to collect valuable insights into how individuals consume and distinguish information, the impact of fake news on their beliefs and decisions, and the measures they think can effectively combat misinformation.

The online survey was made available for participation from September 25$^{th}$ to November 25$^{th}$, 2023. The questionnaire was distributed as widely as possible, and no filters were applied on participants in order to ensure variation in the target audience. Before completing the survey, participants were explicitly notified about the research nature of the survey. To evaluate a participant's existing knowledge without any external influence, the survey intentionally omitted an initial definition of technology. All responses were anonymized, and participants had the option to skip any question as none were obligatory. Participation was entirely voluntary, allowing individuals to exit the survey at their discretion.

To propose the questions for the survey, research results from [81, 82, 37, 25] were used. The selection of primary thematic domains, along with their associated inquiries, and all survey content and the data gathered during this research are openly accessible on https://rb.gy/kbt7lv.

### 4.1.1 Study design

The objective is to investigate the relationship between the five-factor model of personality, individual attitudes, and fake news detection ability. Drawing from the extensive literature review on socio-demographic, behavioral, and intercultural factors that impact the identification of fake news, this study organizes a series of sections [S1, S2,..., S5], each specifically designed to address corresponding Research Questions [S1Q1, S1Q2, ..., S2Q1, ...].

The survey encompassed a diverse range of question formats, including Yes/No, Likert scales, checkboxes, and multiple-choice responses. In questions using a Likert scale, participants were provided with either a 3-point scale ranging from 'agree' to 'disagree' or a 5-point scale spanning from 'strongly disagree' to 'strongly agree' (or 'not at all attentive' to 'extremely attentive', 'extremely disinterested' to 'extremely excited', 'never' to 'always', 'not familiar' to 'very familiar', 'unable' to 'able').

The survey comprised a total of 43 questions, organized into five distinct sections:

- **Demographic Information**: the first section (S1) included six questions [S1Q1-S1Q6] related to age, gender, country, level of education, current occupation, and income level.
- **Personality Traits**: comprising 14 questions [S2Q1-S2Q14], this section (S2) aimed to analyze how the respondent interacts with their environment, including behaviors, thoughts, and emotions.
- **Knowledge and Awareness**: summarized in 6 questions [S3Q1-S3Q6], this section (S3) assessed the respondent's understanding and consciousness of information, facts, or truths about a particular subject or their surroundings.
- **Attitudes Towards News and AI**: the fourth section (S4) contained nine questions [S4Q1-S4Q9] designed to detect the beliefs, perceptions, and stances individuals hold regarding the intersection of artificial intelligence and the consumption or dissemination of news and information.
- **Distinguishing Between Real and Fake News Generated by Artificial Intelligence**: this last section (S5) included six questions [S5Q1-S5Q6] aimed at measuring the respondent's ability to identify and discern the authenticity of news articles and information generated using AI technologies.

When designing a survey, it is imperative to establish a meaningful correlation between the different sections of the questionnaire. This correlation is a key determinant of how respondents are likely to respond to different parts of the survey. For this, aligning and interconnecting the sections of the survey facilitates a smoother flow of information and improves the overall understanding of respondents' views to provide a comprehensive view of respondents' capacity to detect fake news. For example, a section exploring demographic information may influence responses to subsequent sections focusing on opinions or behaviors [83]. Also, the personality traits section affects the decision of surveyors [84].

#### 4.1.2 Data collection

The survey was made available on the Internet and distributed via various channels, including email, Facebook, and other social media platforms. The target audience was diverse, encompassing a wide range of age groups and educational backgrounds, and distributed across different geographical regions worldwide.

The study gathered a total of 103 responses. After excluding incomplete submissions, 83 valid responses were utilized for data analysis. The participant demographics encompassed a diverse range, including 1 individual below 18 years old, 40 participants aged 18 to 29, 33 between 30 and 49, and 8 between 50 and 64. The educational backgrounds of participants varied, spanning from primary education to higher education, and included a mix of unemployed individuals, working professionals, and those currently pursuing studies.

#### 4.1.3 Evaluation method

Survey evaluation entails employing various methods to ensure the validity, reliability, and pertinence of the gathered measures. Among these techniques, structural equation modeling (SEM) emerges as a potent tool [87, 86, 85]. SEM aids in evaluating both convergent and discriminant validity of measures, offering insights into the precision of the survey's measurement instruments. Moreover, SEM enables the modeling of relationships between variables, thereby simplifying the analysis of the intricate structures inherent in surveys. In addition to SEM, other statistical evaluation approaches [12] [13] (mean, median, standard deviation, variance, ...) encompass reliability analysis to gauge the internal consistency of questions, factor analysis to unveil relationships between variables, and prior validation of the questionnaire by experts to ensure conceptual relevance.

SEM not only facilitates the measurement of correlations but also enables the exploration of causal relationships between survey sections. This capability aids in discerning whether one section of the survey significantly influences another. This provides a more deeper understanding of the dynamics between variables. Such insights serve to enhance the interpretation of survey results and offer more comprehensive understanding of the intricate interactions among the various dimensions. These collective methods fortify the validity of the data derived from the survey and make easier interpretation of the results. For the present study, SEM is predominantly employed in assessing the responses to the proposed survey.

### 4.2 New extended real and fake news (ERAF-News) dataset

The literature has introduced datasets for news that cover two or three classes: Real, Fake, and possibly AI-generated from real sources [88]. However, there is no distinction between AI-generated news from real or fake sources.
With the dataset ERAF-News, we aim to differentiate between these two new categories. In fact, distinguishing be-

[12] https://shorturl.at/pwzW8, Last access: 13 October 2023
[13] https://t.ly/fzrV8, Last access: 7 November 2023

tween real and fake news generated by AI offers several advantages, such as: (1) improving credibility by distinguishing between types of AI-generated information; (2) raising public awareness about the presence of AI-generated misinformation by verifying the accuracy of information regardless of its origin; (3) developing automated verification systems to improve algorithms and systems for authenticity and accuracy of online content; (4) informed decision-making based on accurate and correct data; and (5) limiting the spread of false information by identifying misleading content.
These objectives highlight the importance of developing tools and methods to analyze and evaluate AI-generated information. To this end, we propose to create our new dataset ERAF-News.

#### 4.2.1 Building ERAF-News: methodology and process

Describing the methodology for dataset preparation is a pivotal phase in the research on fake news detection. To kick-start the research, the "Fake News Detection Datasets" from Kaggle was used [14]. This dataset encompasses two primary files: "True.csv", for authentic news, and "Fake.csv", containing fake news. These initial files consisted of roughly 23,000 and 21,000 entries, respectively, and included essential columns such as title, text, subject, and date.

The current study introduces a detection framework extending beyond the binary "True" and "Fake" classification. An additional category, "Generated", was incorporated specifically tailored for AI-generated news. This involved extracting data from "True.csv" and "Fake.csv" to create two new sub-datasets named "GAI-X-True.csv" and "GAI-X-Fake.csv" where X denotes the name of the model used for the generation. Consequently, six distinct dataset files were maintained, each representing a unique news category. The objective behind this expanded categorization is to provide a more nuanced and precise foundation for classification.

Figure 2 illustrates the process of obtaining the generated ERAF-News dataset. For the generation process, the choice from the introduced models in section 3.2 was guided by considerations mentioned from [89]. They showed that GPT, BERT, and BART are versatile, capable of both generation and classification tasks, and have demonstrated commendable accuracy.

#### 4.2.2 Generation techniques

The creation of the generated dataset involved a series of generation experiments executed on the Google Colab platform. Pre-trained models were employed to process the original text and produce the corresponding segments of what is now referred to as the "Extended Real And Fake News Dataset" (ERAF-News dataset). This comprehensive dataset offers a valuable resource for researchers and developers looking to advance AI-generated fake news detection. We've made an extract of ERAF-News dataset freely available for unrestricted use on shared drive available at the link https://rb.gy/kbt7lv in order to facilitate collaboration and further research in this critical domain.

From "True.csv" and "Fake.csv", 15000 entries were randomly selected. These entries were used as inputs for generating new classes using the BART model, resulting in "GAI-BART-True.csv" and "GAI-BART-Fake.csv". Likewise, the same set of inputs was used for the GPT-2 model, yielding "GAI-GPT2-True.csv" and "GAI-GPT2-Fake.csv". Concluding the process, upon applying BERT to identical inputs, "GAI-BERT-True.csv" and "GAI-BERT-Fake.csv" are generated. The performance metrics used to evaluate the generation process are discussed in section 5.2.

This dataset was created specifically to be used in the novel classifier architecture, as discussed in the next section.

### 4.3 Proposed dual-stream transformers model (DuSTraMo)

In this section, we present a pioneering strategy to address the issue of news classification, introducing an innovative dual-stream deep learning framework. Given the escalating prevalence of fake news and disinformation, it's imperative to have advanced solutions that can swiftly and precisely determine the accuracy of news articles. We propose an approach based on a dual-stream architecture illustrated in figure 3.

Each stream in the dual-stream model is assigned a pre-trained deep learning model. The outputs of these two streams are subsequently combined through concatenation, enabling the global model to efficiently learn the cross-modal interactions and synergies between the viewpoints of the two models. This fusion of information produces more complete and contextually rich representations, often leading to improved performance in downstream tasks. Fine-tuning with additional dense layers further tailors the entire model to the specific target task, creating a versatile framework adaptable to various classification tasks.
The dual-stream model enhances the learning of complex patterns and relationships, thereby improving the overall accuracy and resilience of classification by enriching feature representation through this architecture [91, 90]. Com-

[14] https://rb.gy/d5mhig, Last access: 25 March 2023

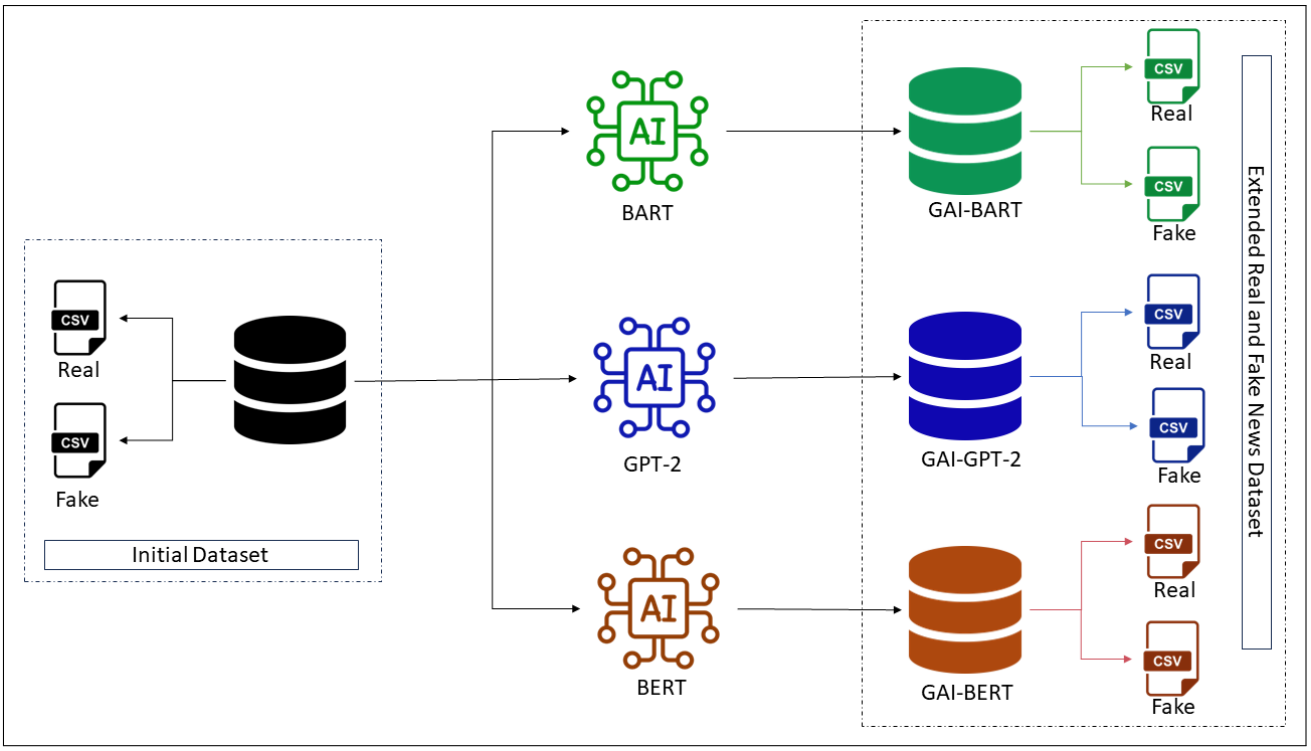



Figure 2: Extended real and fake news (ERAF-News) dataset

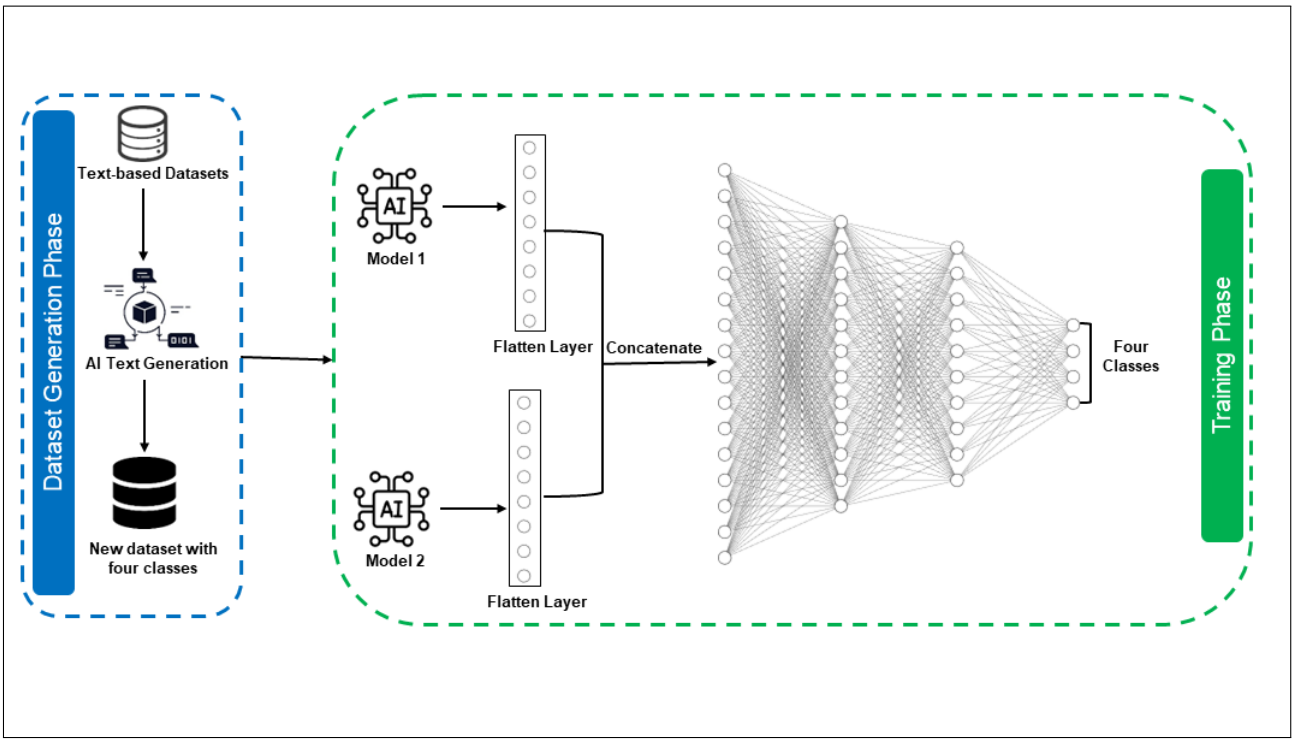



Figure 3: Dual-Stream transformers model (DuSTraMo)

pared to single-model techniques, dual models offer several advantages. Firstly, they perform exceptionally well by combining the strengths of two specialized models, resulting in increased accuracy and improved generalization. Secondly, the synergy between dual models allows for the capture of multiple facets of the data, potentially leading to more accurate predictions. Ultimately, by presenting multiple perspectives on the data, this model has the potential to enhance the explainability of the obtained results [94, 93, 92]

For the classification tasks, the choice among the models cited in section 3.2 was guided by [59, 95] which involved a comparative analysis of the BERT, BART, GPT-2, RoBERTa and DistilBERT models, revealing the competence of these models.

In this paper, we emphasize flexibility in the dual-stream architecture. At any given instance, the first stream could be a GPT-2, BART, or BERT model. The second stream introduces variability, which could consist of the same model as the first or alternatively, DistilBERT or RoBERTa. This yields three distinct configurations:

- The GPT-2-based configuration includes dual-GPT-RoBERTa, dual-GPT-DistilBERT, and dual-GPT-GPT combinations.
- The BART model-based configuration comprises dual-BART-RoBERTa, dual-BART-DistilBERT, and dual-BART-BART.
- The BERT model-based configuration encompasses dual-BERT-RoBERTa, dual-BERT-DistilBERT, and dual-BERT-BERT combinations.

In summary, the three models GPT-2, BART, and BERT serve as constants in the first stream of the dual model due to their roles in the generation process. Additional models, such as RoBERTa and DistilBERT, along with duplicated first stream models, are incorporated into the second stream, showcasing notable classification capabilities as evidenced by previous literature.

To the best of our knowledge, the architecture of a dual-stream classifier has been introduced in the fake news domain only twice: first in [57], which employed simpler (Machine Learning) streams (such as SVM, LSM, RNN, ...), and more recently in [17], which used BERT in both streams. In contrast, the present paper proposes a novel approach by utilizing diverse streams based entirely on transformers.

# 5 Findings

## 5.1 Assessment of survey results

In this section, the study and analysis of the survey results are presented using the SEM methodologies. The survey collected responses from 103 participants, 83 of which were valid, spanning 9 countries around the world. The subsequent analysis is conducted on individual sections of the survey, followed by an examination of the interplay and mutual influences among them.

### 5.1.1 Survey evaluation metrics

The SEM method can include several metrics, depending on its configuration and the specifications of the structural model [96]. Commonly used metrics in SEM include:

- **Chi-square ($\chi^2$):** evaluates the fit of the model to the observed data. A low chi-square indicates a good fit.
- **Degrees of Freedom:** represents the maximum number of logically independent values.

Other metrics commonly used in the fields of statistics, research, and data analysis are:

- **Conditional Demographic Disparity (CDD):** assesses whether a facet exhibits a higher proportion of rejected outcomes in the dataset compared to accepted outcomes.
- **p-value:** indicates the probability that the observed results in statistical analysis are due to chance. In the context of hypothesis testing, a low p-value (typically $< 0.05$) suggests that the results are statistically significant, providing evidence to reject the null hypothesis.
- **Coefficient of Determination ($R^2$):** indicates how well the independent variables can explain the variation in the dependent variable.
- **Completion Rate:** measures the proportion of people who answered all the questions relative to the total number of people invited or contacted. It is an indicator of the survey's engagement and effectiveness.
- **Correlation:** measures the statistical relationship between variables. It can be positive (variables move in the same direction), negative (variables move in opposite directions), or null (no linear relationship). It quantifies the strength and direction of this relationship.

In addition, various metrics can be used to measure different characteristics or aspects namely: frequency, percentage, mean, standard deviation, mean score, etc.

### 5.1.2 Demograhic evaluation

Conducting a demographic assessment in survey research is crucial for gaining a comprehensive understanding of the surveyed sample. Table 3 resumes the participants' data according to the demographic section.

Figure 4 illustrates the distribution of survey participants based on their respective countries.

Conducting a comparative analysis of information regarding "country residence" and "income level," the following outcomes were obtained:

- The Cramer Dependence Coefficient (CDD) registered a value of 35.22%, signifying a substantial correlation between the "country" and "income level" variables. This CDD value points to a notable association, implying that the income level is closely linked with the respondent's country of residence.

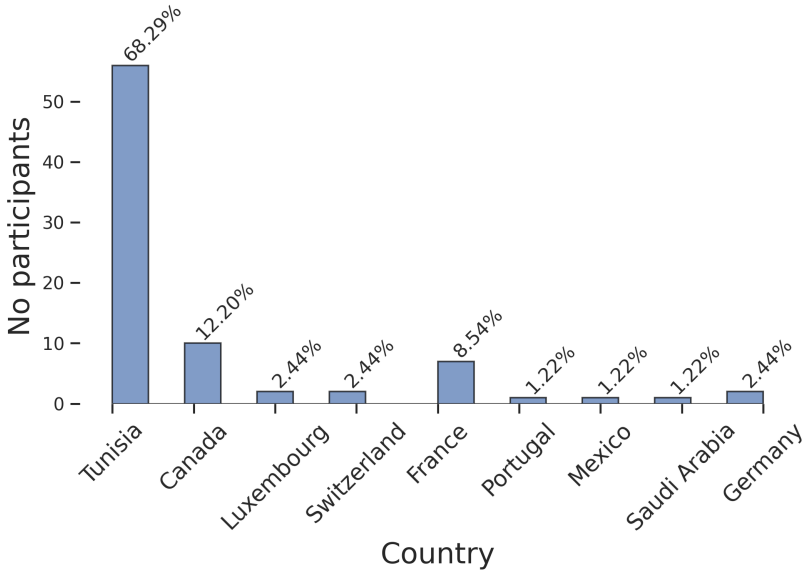



Figure 4: Countries participation histogram

Table 3: Participants demographic data

| Variable | Values | N (%) |
|---|---|---|
| Gender | Male | 44 (42.7%) |
| | Female | 59 (57.3%) |
| Age | Under 18 | 1 (1%) |
| | 18-29 | 50 (48.5%) |
| | 30-49 | 42 (40.8%) |
| | 50-64 | 10 (9.7%) |
| | 65 or older | 0 |
| Current occupation | Student | 45 (44.1%) |
| | Worker | 56 (54.9%) |
| | Unemployed | 1 (1%) |
| Education Level | Secondary Education | 2 (2%) |
| | Bachelor's degree | 13 (13%) |
| | High school or equivalent | 16 (16%) |
| | Master's degree | 46 (46%) |
| | Doctoral degree | 20 (20%) |
| | Prefer not to answer | 3 (3%) |
| Income Level | Low | 20 (19.6%) |
| | Medium | 45 (44.1%) |
| | High | 13 (12.7%) |
| | Prefer not to answer | 24 (23.5%) |

– The Chi-square test statistic yielded a relatively high result of 30.52%, measuring the disparity between observed data and anticipated values under the null hypothesis, which assumes independence between the variables "country" and "income level." This outcome underscores a significant association between the two variables.

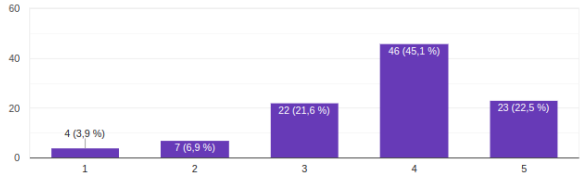
– The associated p-value, measuring the likelihood of obtaining results as extreme as those observed, was calculated as 16.81% in the context of the Chi-square test. Although this value exceeds the commonly used significance threshold of 5%, indicating insufficient evidence to reject the null hypothesis of independence, it is noteworthy that the p-value is not extremely high, suggesting a tendency towards association despite non-rejection.

– A total of 24 degrees of freedom were observed, a parameter contingent on the size of the contingency table. This substantial degree of freedom, reflective of the relatively large dataset, reinforces the statistical robustness of the analysis. In summary, these findings underscore a moderate to a strong association between "country of residence" and "income level", and they affirm the adequacy and reliability of the analyzed sample comprising 83 valid survey responses.

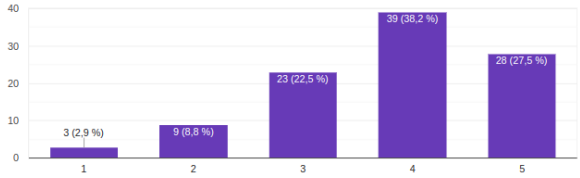
Similarly, an analysis of "age" and "current occupation" reveals a strong and statistically significant association, as evidenced by a CDD of 47.80%, a high chi-square test value of 37.47%, and an extremely low p-value 1.42e-06 with an interaction degree of freedom of 6.

### 5.1.3 Personality traits evaluation

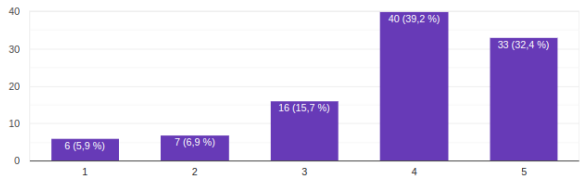
The second section (named S2) of the questionnaire delves into personality traits. As illustrated in figure 5, it's evident that a significant portion of individuals who value organization and tidiness (44.3%) also tend to engage in advance planning (38.1%) and exhibit enthusiasm for exploring new experiences (39.2%). These characteristics notably impact an individual's disposition during conversations, with 58.3% reporting a high level of comfort.



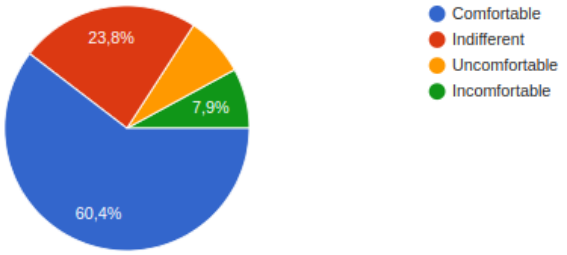

(a) Having an organized workspace.




(b) Tend to plan activities in advance.




(c) Ability to try new experiences.




(d) Feeling in initiating conversations.

Figure 5: Global personality traits analysis

Figure 6 summarises an assessment using the Big Five method produced mean values for Openness in [0-2], Conscientiousness in [0-3], Extroversion in [0-3], Agreeableness in [0-2] and Neuroticism in [0-2]. Specifically, the obtained results revealed mean and standard deviations for the Big five personality factors as follows: openness (M = 1.21; SD = 0.52), extroversion (M = 1.21; SD = 0.51), agreeableness (M = 1.19; SD = 0.53), neuroticism (M = 0.75; SD = 0.40), and conscientiousness (M = 1.64; SD = 0.53).

#### 5.1.4 Knowledge and awareness evaluation

Analyzing the results of the third section (S3), the survey participants indicated that many of them rely on international TV for their news, and several of them do not actively check its accuracy. However, most of them have limited knowledge about concepts such as fake news, and AI, and are unfamiliar with the term ”ChatGPT”. This is due to a lack of awareness about GAI.

Figure 7 illustrates the breakdown of survey participants' responses regarding their level of knowledge and awareness.

As outlined in section 4.1.1, the third section of the survey (S3) comprises eight questions (S3Q1-S3Q8). The distribution of responses has been assessed for each question. The responses have undergone pre-processing and normalization, resulting in values within the 0 to 1 range. The initial histogram illustrates the distribution of responses to the first question (S3Q1), with the predominant choice being the second answer. Similarly, for the histogram corresponding to S3Q2, the majority of participants selected the third answer, and this pattern persists across the subsequent histograms.

#### 5.1.5 Attitudes towards news and AI evaluation

This section explores respondents' perspectives on fake news, including their ability to detect it, their capacity for verification, and their potential for guarding against unintended consequences of AI-generated fake news. Following data pre-processing, figure 8 summarizes the calculated attitude scores.

As this section comprises nine questions with scores that can range from 0 to 9, it is noteworthy that 54.90% of respondents attained scores between 4 and 6 (figure 8a). This observation suggests that a significant portion of the participants exhibit psychological equilibrium in their attitudes toward fake news. Furthermore, a p-value was calculated to assess the relationship between the respondent's country of residence and their attitude scores. The p-value, approximately 7.82%, suggests that there are statistically significant differences in attitude scores across the various countries. Even though 57.31% of the respondents rely on the Internet for 75% of their information, it is noteworthy that 47.56% of them expressed uncertainty about their ability to discern fake news from headlines alone (figure 8b).

It can be additionally observed that the practice of utilizing the Internet to acquire information is associated with the capacity to discern fake news. Specifically, as shown in figure 9, the linear correlation matrix between the variables of this section and its general score reveals that question S4Q2 (”about the use of the Internet as a source of information”), exerts the most significant variance on this score, accounting for 50%.

#### 5.1.6 Distinguishing between real and fake news generated by AI evaluation

The fifth section (S5) contains six questions [S5Q1-S5Q6], with score means that ranged from 0 to 6. The survey results

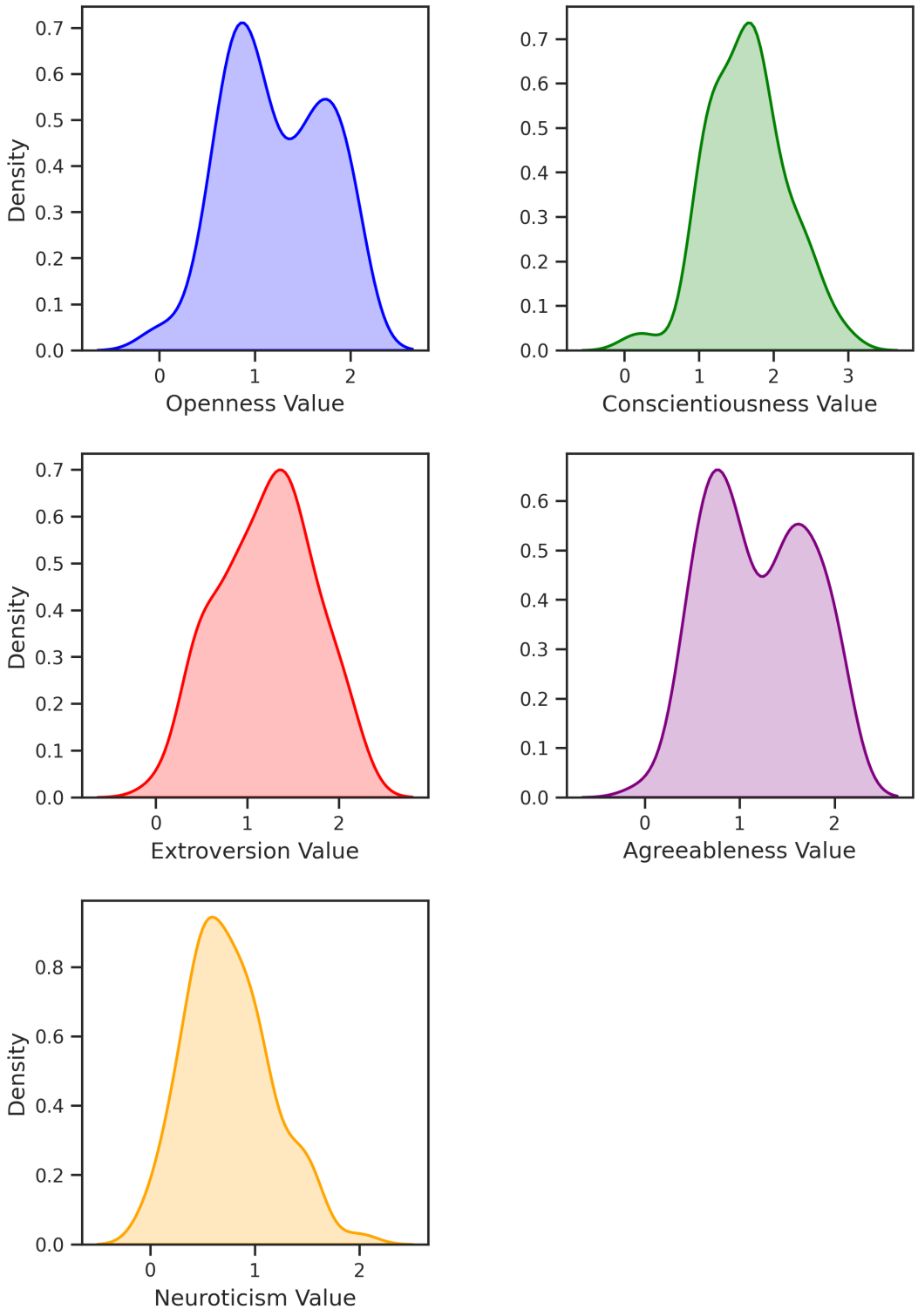



Figure 6: Assessment curves of participants' personality traits

showed an average score of 2.73, indicating that they are only able to distinguish between fake and real news to a degree of 40.19%.

Additionally, a computation of the correlation matrix, as depicted in figure 10, reveals the highest correlation at approximately 45%. Such a correlation is typically categorized as moderate. This suggests that there exists an average relationship between questions S5Q1 and S5Q2.

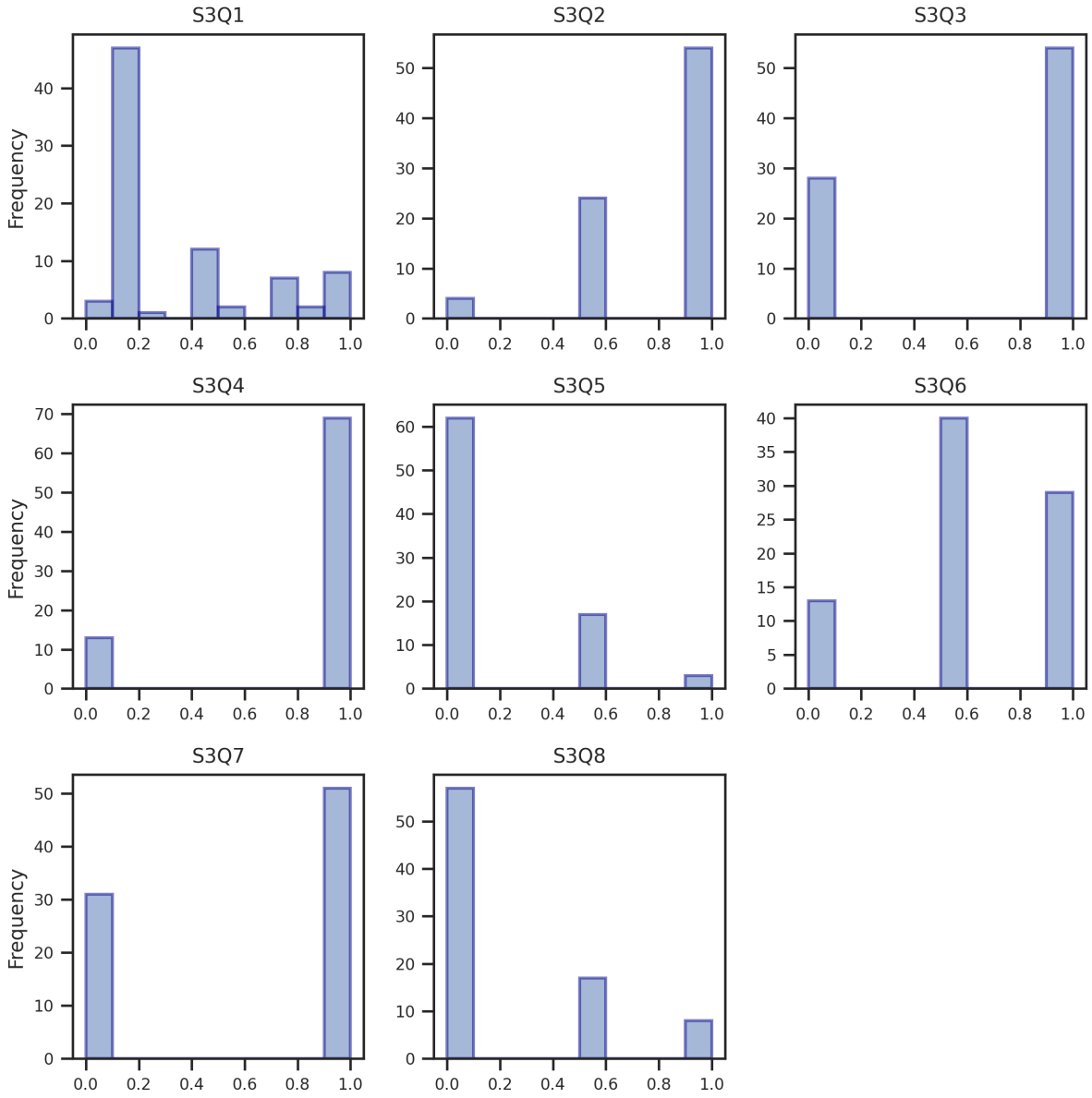



Figure 7: Distribution of knowledge level and awareness

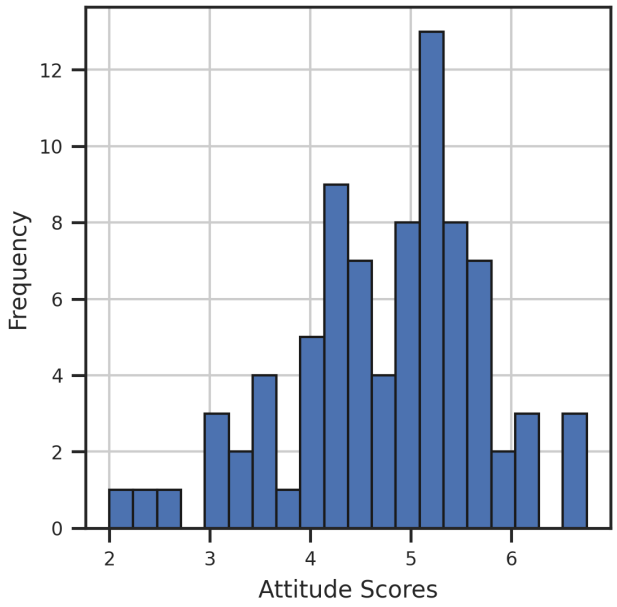
### 5.1.7 Survey summary

A total of 103 survey participants engaged in the study through an internationally published online questionnaire over 3 months. The uniqueness of this investigation resides in its examination of individuals' capacity to discern fake news based on the influence of demographic characteristics and personality traits according to the Big five personality model (OCEAN). In figure 11, a correlation analysis was performed using the five personality traits from the Big Five section (S2) to demonstrate their relationship on the sections S3 ('Knowledge and Awareness'), S4 ('Attitudes Towards News and AI'), and S5 ('Distinguishing Between Real and Fake News Generated by AI').

The findings reveal that:

- Section S3 ("Knowledge and Awareness") is most notably correlated with openness (61%) and extroversion (49%).
- Conscientiousness exerts a 43% correlation on Section S4 ("Attitudes Towards News and AI").
- Conversely, Section S5 ("Distinguishing Between Real and Fake News Generated by AI") is primarily explained by the agreeableness characteristic, contributing to 68% of its influence."

The results indicate that the percentages of intersecting influences are not exceedingly high, hovering around 50%. While the evaluations did underscore distinctions within




(a) Number of responses by attitude score

| S4Q8 / S4Q2 | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 1 | 4 |
| 2 | 4 | 2 | 6 | 1 | 1 | 14 |
| 3 | 5 | 2 | 8 | 1 | 1 | 17 |
| 4 | 15 | 3 | 10 | 2 | 0 | 30 |
| 5 | 7 | 0 | 7 | 1 | 2 | 17 |
| Total | 31 | 8 | 32 | 6 | 5 | 82 |

(b) Cross table between habit of using the internet to gather information and ability to identify fake news

Figure 8: Evaluation of the ”Attitudes Towards News and AI Evaluation”

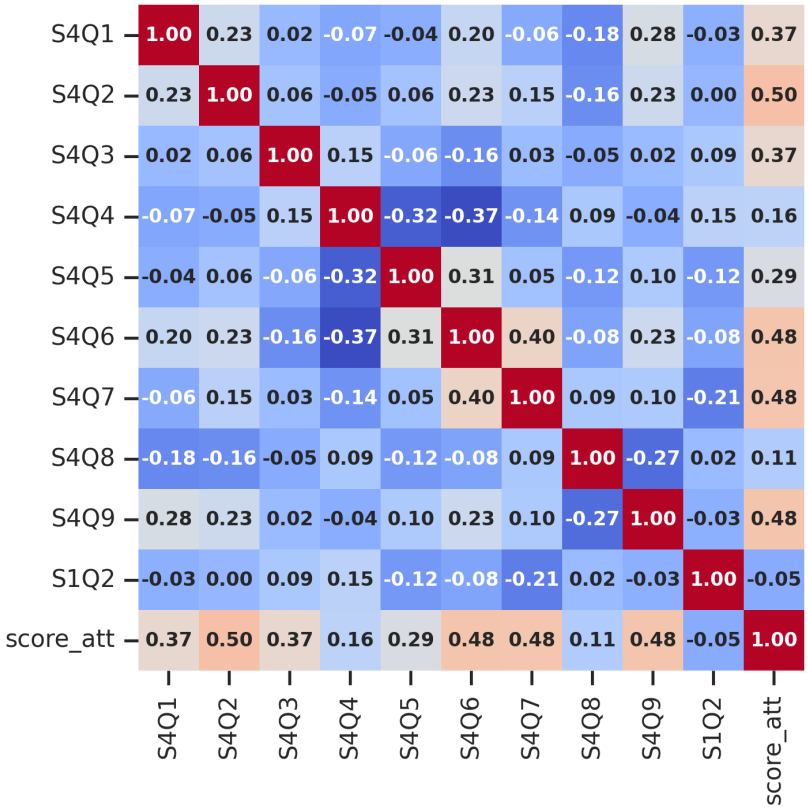



Figure 9: Correlation of Internet use on the attitude of responders

survey results, they nonetheless proved inadequate to attain a robust level of performance. Consequently, the study grapples with certain limitations. Firstly, the study's outcomes were derived from a relatively modest cohort of 83 users possessing personalities of moderately acceptable openness, conscientiousness, and extroversion.

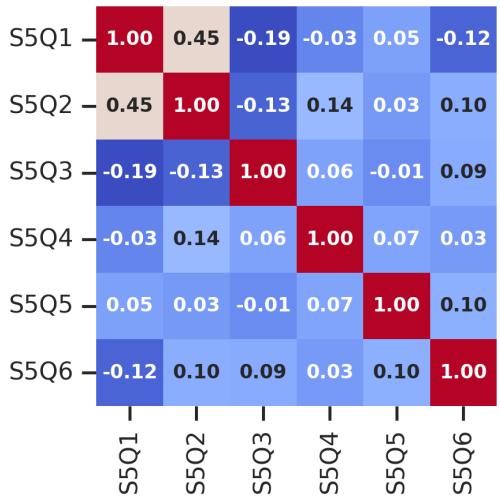



Figure 10: Correlation matrix for distinguishing between GAI real and fake news

#### 5.1.8 Discussion and limits of surveying method

The outcomes of the survey indicate that only a small proportion of participants potentially can discern between authentic and fabricated news, with the majority appearing to lack this skill. This finding accentuates the necessity for an automated method to detect fake news, aimed at assisting users in identifying misinformation.

In this context, the application of AI serves as a crucial tool to enhance the effectiveness of these detection strategies. To ensure improvement in results through AI techniques, a study was conducted on survey questions. Specifically, we tested GPT, BART, and BERT models to predict responses. Table 4 showcases the results of the predictions made.

Table 4: Assessment of survey question prediction

| | GPT | BART | BERT | DistilBERT | RoBERTa |
|---|---|---|---|---|---|
| GAI-GPT | **0.82** | 0.71 | 0.77 | 0.50 | 0.25 |
| GAI-BART | 0.45 | 0.62 | 0.68 | 0.33 | 0.17 |
| GAI-BERT | 0.67 | 0.33 | 0.50 | 0.50 | 0.50 |

It was noted in section 5.1.6 that survey respondents can correctly answer questions from Section S5 with a percentage of 40.19%. However, with the integration of AI prediction, it has been demonstrated that we can achieve even a percentage of 82% of accuracy.

As we progress with this research, we strongly advocate for the integration of AI, which promises to introduce new, more robust dimensions in the realm of fake news detection. This forward-thinking approach is expected to surmount existing challenges by harnessing the power of machine learning and sophisticated data analysis, thereby revolutionizing the current landscape of misinformation identification.

The next section discusses an automated system grounded in artificial intelligence to surmount identified challenges and markedly enhance the performance of incorrect information detection.

### 5.2 Assessment of data generation quality

As previously mentioned, the ERAF-News dataset serves as the data source, comprising three distinct generated setsusing pre-trained models: BART, BERT, and GPT-2. The generated output deserves to be evaluated to measure the degrees of similarity and fluency of the generated text compared to the original text.

In the following, a list of metrics used are defined, then an evaluation of the generation quality of the ERAF-News dataset is presented.

#### 5.2.1 Generation evaluation metrics

Metrics play a crucial role in evaluating results following text generation. The most commonly used metrics in the generation context are:

1. **BLEU score** (Bilingual Evaluation Understudy) focused on precision, initially used for translation but can be leveraged for generation evaluation [15]. BLEU score typically ranges from 0 to 1, where a score closer to 1 indicates a higher similarity between the generated text and the reference texts.

2. **ROUGE score** (Recall-Oriented Understudy for Gisting Evaluation) emphasizes recall, is Based on n-gram overlap [16]. Higher ROUGE scores indicate better performance in terms of matching the generated text to the reference texts.

3. **BERT score** is a metric designed to evaluate machine translation. It calculates a similarity score between each token in the candidate sentence and each token in the reference sentence. This is achieved by utilizing contextual embeddings from pre-trained BERT models and comparing words in candidate and reference sentences using cosine similarity. Additionally, BERT score provides valuable insights for the evaluation of diverse language generation tasks [97].

4. **BLEURT score** (Bilingual Evaluation Understudy with Representations from Transformers) is a metric designed for evaluating the quality of machine-generated text [98]. It focuses on evaluating the fluency and adequacy of generated text.

[15] http://tinyurl.com/2p96r5bx, Last access: 23 August 2023
[16] https://n9.cl/17pto, Last access: 23 August 2023

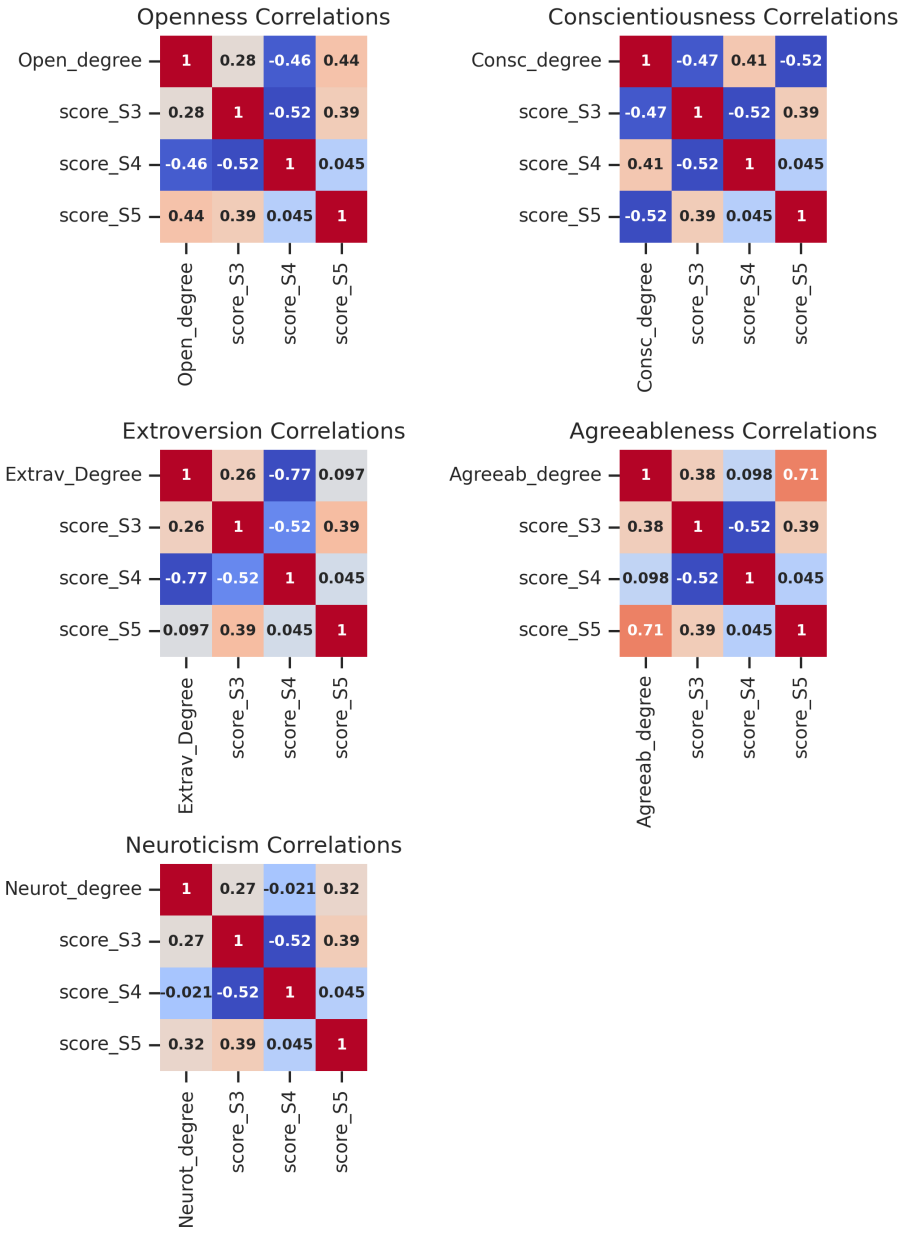



Figure 11: Influence of big five traits on other sections

### 5.2.2 Generation results

In Table 5, the evaluation of the generated dataset was conducted using various metrics and compared to prior studies. This assessment aimed to measure how closely the generated dataset resembles the original one.

The results of BLEU and ROUGE scores varied in the literature between [0.089-0.5] and [0.1-1.2] respectively. The multi-class classification carried out on ERAF-News exceeds these results and shows superior performances of the order of [0.203-0.256] and [0.366 - 0.916] respectively. In terms of BERT and BLEURT scores which varied between [0.2-0.6] and [0.3-0.7] respectively in the literature, reach overall scores higher than 0.91 and 0.85 respectively when carried out on ERAF-News dataset. As a conclusion, among the evaluated metrics, BART consistently emerges as the top-performing model, excelling in BLEU, ROUGE, BERT scores, and BLEURT scores. Its comprehensive success across multiple evaluation criteria positions BART as the most robust and effective model for text generation tasks. Additionally, BERT demonstrates strong performance across various metrics, showcasing its versatility in generating text that aligns closely with reference material.

## 5.3 Assessment of proposed method

This section provides a detailed exploration of the experiments conducted and a thorough analysis of the results ob-

Table 5: Comparative study of generation results

| Year | Ref. | Dataset | Model | BLEU | Rouge | BERT score | BLEURT score |
|---|---|---|---|---|---|---|---|
| 2019 | [89] | WMT'16 ELI5 | BART | 0.379 | 1.2 | - | - |
| 2020 | [97] | WMT18 | BERT | 0.527 | 0.536 | 0.693 | - |
| | [98] | WMT17-19 | BERT | 0.2-0.3 | - | 0.25-0.45 | 0.3-0.57 |
| 2021 | [99] | CovidDialog | GPT2 | 0.094 | - | - | - |
| | | | BART | 0.089 | - | - | - |
| 2022 | [100] | WMT19 | BART | 0.206 | 0.165 | 0.317 | 0.325 |
| | [101] | WebText CNNDM | NLP models | - | 0.286 | 0.332 | 0.716 |
| 2023 | current study | ERAF-News | GAI-GPT | 0.203 | 0.366 | 0.917 | 0.853 |
| | | | GAI-BART | 0.216 | 0.383 | 0.957 | 0.983 |
| | | | GAI-BERT | **0.256** | **0.916** | **0.918** | **0.929** |

tained from the proposed method. The evaluation begins by assessing the individual performance of the selected models, followed by evaluating the dual models using various combinations.

#### 5.3.1 Experimental setup

Single model executions were conducted on Kaggle, using a GPU T4x2 accelerator. On the other hand, for the dual model executions, Google Colab Pro and Kaggle platforms were used, leveraging GPU processing.

Several important parameters were used during the model's execution phase to guarantee efficient training and assessment. The dataset is divided into a training set comprising 80% of the data, and a test set comprising the remaining 20% of the data. The training process utilized a batch size of 2 for both single and dual models, with model weights adjusted over 2 epochs for each. Additionally, training performance was monitored and hyper-parameter adjustments are made using a validation split of 20% of the training set.

#### 5.3.2 Used metrics

Evaluating the performance of various models includes the examination of detection metrics such as accuracy, precision, recall, and F1-score [103, 102]:

- **Accuracy:** the percentage of all correctly identified data points and the percentage of all correctly predicted data points (both positive and negative) are the same.
- **Precision:** the ratio of true positives to all real positive instances is the same as the ratio of true positives to all positive predictions.
- **Recall:** the proportion of true positive predictions among all actual positive cases is the same as the proportion of true positives among all positive predictions.
- **F1 score:** F1 score is calculated by taking the harmonic mean of the model's precision and recall.

These metrics are commonly used to measure the performance of machine and deep learning models, particularly in classification tasks.

#### 5.3.3 Results of single models multi-classification

Table 6 displays the performance metrics of various single models multi-classification on different datasets: GAI-GPT2, GAI-BERT, and GAI-BART along with a comparison to models documented in the literature.

In the literature, previous works with single transformer models for classification predominantly utilized models such as RoBERTa, BERT, and XLNet. Their evaluations were primarily based on Accuracy and F1-score, ranging between 0.7 and 0.9. In contrast, the present study undertakes the RoBERTa and BERT models by incorporating additional models such as GPT-2, BART and DistilBERT and includes measures of precision and recall scores.

Multi-class classification with RoBERTa, regardless of the dataset, emerged with the lowest performance, with an accuracy of 0.2 and all other metrics being null, demonstrating its incapacity to multi-classify any type of generation. Similarly, BART showed very weak metrics, around 0.2, indicating its inability to perform multi-classification for any type of generation. In contrast, the BERT and DistilBERT models achieved perfect metrics, indicating that they made major predictions correctly. GPT-2 obtained equivalent values for all metrics, around 0.7, reflecting relatively consistent performance. Finally, GPT-2 demonstrates a balance between the values of its metrics but with lower performance compared to BERT and DistilBERT.

In summary, DistilBERT demonstrates excellent classification capabilities of all datasets, making it a promising choice for applications requiring high precision in class predictions.

Table 6: Comparative study of single models results

| Year | Ref. | Dataset Used | | Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|---|---|---|
| 2021 | [104] | PolitiFact | | RoBERTa | 0.825 | - | - | 0.805 |
| | | | | BERT | 0.88 | - | - | 0.87 |
| | | | | XLNet | 0.895 | - | - | 0.90 |
| | | GossipCop | | RoBERTa | 0.803 | - | - | 0.807 |
| | | | | BERT | 0.85 | - | - | 0.79 |
| | | | | XLNet | 0.855 | - | - | 0.78 |
| 2022 | [95] | LIAR | | BERT | 0.77 | - | - | - |
| 2023 | [105] | real-world | | BERT | 0.8843 | 0.8937 | 0.8756 | 0.8846 |
| | [72] | In-House | | BERT | 0.919 | - | - | - |
| | | | | RoBERTa | 0.961 | - | - | - |
| | | TweepFake | | BERT | 0.891 | - | - | - |
| | | | | RoBERTa | 0.911 | - | - | - |
| | current study | ERAF-News | GAI-GPT2 | RoBERTa | 0.2530 | 0.0000 | 0.0000 | 0.0000 |
| | | | | BERT | **0.9787** | **0.9787** | **0.9787** | **0.9770** |
| | | | | GPT2 | 0.7245 | 0.7241 | 0.7241 | 0.6348 |
| | | | | DistilBERT | **0.9845** | **0.9845** | **0.9845** | **0.9833** |
| | | | | BART | 0.2463 | 0.2460 | 0.2460 | 0.0973 |
| | | | GAI-BART | RoBERTa | 0.2467 | 0.0000 | 0.0000 | 0.0000 |
| | | | | BERT | **0.9737** | **0.9722** | **0.9721** | **0.9760** |
| | | | | GPT2 | 0.6065 | 0.7460 | 0.4262 | 0.4729 |
| | | | | DistilBERT | **0.9643** | **0.9647** | **0.9638** | **0.9604** |
| | | | | BART | 0.2468 | 0.2468 | 0.2468 | 0.0977 |
| | | | GAI-BERT | RoBERTa | 0.2412 | 0.0000 | 0.0000 | 0.0000 |
| | | | | BERT | **0.9775** | **0.9775** | **0.9775** | **0.9759** |
| | | | | GPT2 | 0.8398 | 0.8403 | 0.8388 | 0.8287 |
| | | | | DistilBERT | **0.9720** | **0.9720** | **0.9720** | **0.9694** |
| | | | | BART | 0.2538 | 0.2533 | 0.2533 | 0.0993 |

#### 5.3.4 Single models discussion

Performance of models varies significantly across datasets, and each one exhibits distinct strengths and weaknesses. DistilBERT and BERT consistently delivered the best results across datasets, showcasing high precision, recall, and F1-scores. On the other hand, BART RoBERTa demonstrated inferior performance with very low Accuracy. GPT-2 model showed reasonably good performance, although with variations across datasets. These findings highlight the nuanced performance characteristics of each model, emphasizing the importance of considering both effectiveness and efficiency in choosing a model for specific applications.

#### 5.3.5 Results of dual models multi-classification

Table 7 offers a comprehensive breakdown of the performance metrics for each dual model across various sub-datasets. The table displays four key evaluation metrics: accuracy, precision, recall, and F1 score.

The literature on dual models for fake news classification is relatively limited. Previous works, as mentioned in [57], incorporated dual-ML models to achieve accuracy ranging between 0.92 and 0.95. The most recent research [17] utilized a dual-BERT-BERT model, achieving an accuracy of 0.85.

Applying DuSTraMo on ERAF-News yielded results in 2 classes: (1) some dual models' highly performing outcomes exceeding 0.95 of accuracy, surpassing the literature's benchmarks; and (2) other dual models showing instability in their results. Sometimes, they provide extremely poor results or very good ones. And at other times they simply did not work at all, especially with dual-BART-BART (indicated as NE for non-executable). In addition, it was very challenging to execute due to their enormous dimension and memory requirements.

For more details, the GPT dual model applied to the GAI-GPT2-generated dataset consistently achieved the remarkably high accuracy score of 0.9999. This observation underscores the model's exceptional performance when applied to datasets intended for GPT2 generation. Conversely, for the GAI-BART dataset, the dual-GPT2-DistilBERT model outperformed other model combinations with an accuracy of 0.9885. Finaly, for the GAI-BERT generated dataset, the BERT-RoBERTa dual model outperformed other model combinations with an accuracy of 0.9881.

#### 5.3.6 Dual models discussion

Present research meticulously explores a range of configurations, employing diverse models to ascertain their ef-

Table 7: Results of DuSTraMo models

| Year | Ref. | Dataset Used | | Dual Models | | Acc. | Prec. | Recall | F1 score |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Model 1 | Model 2 | | | | |
| 2022 | [57] | CONSTRAINT shared task-2021 | | BERT-XLNet-ELMo | | 0.93 | - | - | 0.925 |
| | | | | LSTM-BiLSTM GRU-BiGRU | | 0.92 | - | - | 0.925 |
| | | | | LR-SVM-RF-KNN BERT | | 0.95 | - | - | 0.95 |
| 2023 | [17] | FakeNewsNet | | BERT | BERT | 0.854 | 0.756 | 0.555 | 0.640 |
| 2024 | Current study | ERAF-News | GAI-GPT2 | GPT2 | GPT2 | **0.9999** | **0.9999** | **0.9999** | **0.9998** |
| | | | | | RoBERTa | 0.9995 | 0.9995 | 0.9995 | 0.9995 |
| | | | | | DistilBERT | 0.9998 | 0.9998 | 0.9998 | 0.9997 |
| | | | | BART | BART | NE | NE | NE | NE |
| | | | | | RoBERTa | 0.9973 | 0.9973 | 0.9973 | 0.9971 |
| | | | | | DistilBERT | 0.9977 | 0.9977 | 0.9977 | 0.9977 |
| | | | | BERT | BERT | 0.3239 | 0.9208 | 0.0729 | 0.1044 |
| | | | | | RoBERTa | 0.9670 | 0.9670 | 0.9668 | 0.9651 |
| | | | | | DistilBERT | 0.9833 | 0.9833 | 0.9833 | 0.9826 |
| | | | GAI-BART | GPT2 | GPT2 | 0.9810 | 0.9810 | 0.9810 | 0.9793 |
| | | | | | RoBERTa | 0.9622 | 0.9624 | 0.9619 | 0.9603 |
| | | | | | DistilBERT | **0.9885** | **0.9885** | **0.9885** | **0.9885** |
| | | | | BART | BART | 0.2500 | 0.2500 | 0.2500 | 0.0985 |
| | | | | | RoBERTa | 0.2485 | 0.2485 | 0.2485 | 0.0982 |
| | | | | | DistilBERT | 0.9563 | 0.9563 | 0.9563 | 0.9540 |
| | | | | BERT | BERT | 0.9753 | 0.9771 | 0.9749 | 0.9746 |
| | | | | | RoBERTa | 0.9851 | 0.9851 | 0.9851 | 0.9845 |
| | | | | | DistilBERT | 0.9758 | 0.9759 | 0.9757 | 0.9748 |
| | | | GAI-BERT | GPT2 | GPT2 | 0.9866 | 0.9883 | 0.9845 | 0.9856 |
| | | | | | RoBERTa | 0.9675 | 0.9678 | 0.9671 | 0.9645 |
| | | | | | DistilBERT | 0.8275 | 0.8652 | 0.8071 | 0.7954 |
| | | | | BART | BART | 0.9735 | 0.9756 | 0.9717 | 0.9729 |
| | | | | | RoBERTa | 0.2515 | 0.2515 | 0.2515 | 0.0992 |
| | | | | | DistilBERT | 0.9710 | 0.9707 | 0.9702 | 0.9684 |
| | | | | BERT | BERT | 0.9833 | 0.9833 | 0.9833 | 0.9821 |
| | | | | | RoBERTa | **0.9881** | **0.9882** | **0.9880** | **0.9873** |
| | | | | | DistilBERT | 0.9868 | 0.9868 | 0.9864 | 0.9854 |

ficacy. Notably, it experiments with GPT-2, BART and BERT models as the primary stream within a dual model structure, coupled with various models in the secondary stream. The empirical results, however, revealed a marked under-performance when juxtaposed with setups where BART served as one of two streams. This recurrent pattern of subpar results, particularly evident in datasets synthesized by both GAI-GPT2 and GAI-BART, points sometimes to a potential inadequacy of the BART model within the context of current experimental framework. The consistently low accuracy scores associated with the BART model underscore its limitations for the tasks and datasets under consideration. This critical insight necessitates a more rigorous and nuanced approach in the selection and application of models for these specific types of computational tasks.

Additionally, it has been observed that dual models incorporating GPT2-X demonstrate a consistent and notable superiority in performance over their BART-X and BERT-X counterparts. This indicates a distinct advantage of GPT2 in the realm of text classification, especially concerning texts generated by both GPT2, BERT and BART, in contrast to the results achieved with BART and BERT models. Such findings raise critical considerations regarding the efficacy and applicability of these models in specific text classification scenarios.

Furthermore, it is imperative to address the influence of dataset characteristics on model performance. This research reveals that the dual model configuration GPT2-X

showed good performance for all generated datasets. This observation may suggest that the BART model possesses a superior capability in the context of data generation tasks. Such a differential impact underscores the importance of dataset selection and its consequential effect on the performance metrics of various model configurations.

Finally, the model associated with the second stream (which varied between RoBERTa, DistilBERT, and a duplicate of the first stream) significantly influenced the performance of the primary model in the first stream. Indeed, RoBERTa and DistilBERT slightly degraded the performance of the other models but remain robust when applied to GAI-GPT2 and GAI-BERT. However, when applied to GAI-BART, the performance degradation is more pronounced. This could be interpreted as poor generation by BART, whereas the generation quality by GPT-2 and BERT is much better than that generated by BART.

#### 5.3.7 Single-stream vs. dual-stream models

To clearly demonstrate the effectiveness of the proposed dual-stream models compared to existing single-stream approaches, we conducted a detailed statistical analysis of the performance metrics (accuracy, precision, recall, and F1 score) across various datasets. Our findings show that dual-stream models consistently outperform single-stream models, as evidenced by significant improvements in all performance metrics. We performed paired t-tests which revealed statistically significant p-values (less to 0.05) for accuracy, precision, recall, and F1 score. Additionally, we calculated 95% confidence intervals for these metrics, demonstrating the reliability and robustness of our results. The dual-stream models not only achieved higher mean performance but also exhibited lower variability, suggesting more consistent results across different datasets. These findings highlight the significant potential of dual-stream architectures in improving the detection and classification of fake news.

## 6 Conclusion

The contributions of this research are threefold. The initial segment delved into surveying the influence of fake news on Internet users and their ability to discern it. Despite employing both Big five criteria and the SEM method, the survey revealed consistently low rates, indicating that information consumers struggle to detect generated fake content, scoring an average of 2.73 within the [0.6] interval. This underscores the elevated performance of fake news generation, rendering it challenging for Internet users to identify them in the majority of cases.

The second phase centered around generating a novel fake news dataset encompassing four types: fake, real, GAI-fake, and GAI-real. Leveraging various standard generators, performance evaluation highlighted BERT's unparalleled efficacy, showcasing impressive metrics with ROUGE (91.6%) and BERT-Score (91.8%).

Conclusively, the third segment demonstrated that our DuSTraMo model classifying the four ERAF-News classes significantly improved the detection performance ensuring an accuracy of 99.12%. Integrating our four-category classification model into existing fake news detection systems improves their accuracy and effectiveness while recognizing nuanced information to better combat misinformation.

While our four-category classification model offers many advantages, it is essential to: (1) control the complexity of the model when adding additional news categories, and (2) consider the ability of DuSTraMo to generalize effectively across languages using Transfer Learning and Federated Learning.

Such research could be considered a starting point towards the future of multi-classes fake news detection, with the possibility of utilizing additional models for comparative purposes.

## Declarations

- Declaration of generative AI and AI-assisted technologies in the writing process : During the preparation of this work, sometimes the authors used ChatGPT in order to improve writing texts. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.
- Funding: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- Conflict of interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: A shared drive is available at the link https://rb.gy/kbt7lv.
- Code availability: Not applicable
- Methods: For the writing of the paper, chatGPT was sometimes used.
- Authors' contributions: In Table 8

Table 8: Authors' contributions

| Authors | Supervision | Bibliographic search | Coding | Discussion results | Writing paper | Content paper |
|---|---|---|---|---|---|---|
| **Hounaida Moalla** | | X | X | X | X | X |
| **Hana Abid** | | X | | | X | X |
| **Dorsaf Sallami** | | X | X | X | X | X |
| **Esma Aïmeur** | X | | | | | X |
| **Bassem Ben Hamed** | X | | | | | X |

# References


[1] Velichety, S., & Shrivastava, U. (2022). Quantifying the impacts of online fake news on the equity value of social media platforms–Evidence from Twitter. International Journal of Information Management, 64, 102474. https://doi.org/10.1016/j.ijinfomgt.2022.102474

[2] Gupta, M., Dennehy, D., Parra, C. M., Mäntymäki, M., & Dwivedi, Y. K. (2023). Fake news believability: The effects of political beliefs and espoused cultural values. Information & Management, 60(2), 103745. https://doi.org/10.1016/j.im.2022.103745

[3] Choi, J., & Lee, J. K. (2022). Confusing effects of fake news on clarity of political information in the social media environment. Journalism Practice, 16(10), 2147-2165. https://doi.org/10.1080/17512786.2021.1903971

[4] Loos, E., & Ivan, L. (2022). Fighting Fake News: A Generational Approach (p. 172). MDPI-Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/soc12020057

[5] Baptista, J.P.; Gradim, A. (2022). A Working Definition of Fake News. Encyclopedia 2022, 2, 632-645. https://doi.org/10.3390/encyclopedia2010043

[6] Leeder, C. (2019). How college students evaluate and share "fake news" stories. Library & Information Science Research, 41(3), 100967. https://doi.org/10.1016/j.lisr.2019.100967

[7] Abu Arqoub, O., Abdulateef Elega, A., Efe Özad, B., Dwikat, H., & Adedamola Oloyede, F. (2022). Mapping the scholarship of fake news research: A systematic review. Journalism Practice, 16(1), 56-86. https://doi.org/10.1080/17512786.2020.1805791

[8] Sallami, D. (2022). Personalized fake news aware recommendation system. https://doi.org/1866/27492

[9] Amri, S., Sallami, D., & Aïmeur, E. (2021). Exmulf: an explainable multimodal content-based fake news detection system. In International Symposium on Foundations and Practice of Security (pp. 177-187). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-08147-7_12

[10] Sallami, D., Ben Salem, R., & Aïmeur, E. (2023). Trust-based Recommender System for Fake News Mitigation. In Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (pp. 104-109). https://doi.org/10.1145/3563359.3597395

[11] Kondamudi, M. R., Sahoo, S. R., Chouhan, L., & Yadav, N. (2023). A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. Journal of King Saud University-Computer and Information Sciences, 35(6), 101571. https://doi.org/10.1016/j.jksuci.2023.101571

[12] Gao, Y., González, V. A., & Yiu, T. W. (2020). Exploring the relationship between construction workers' personality traits and safety behavior. Journal of construction engineering and management, 146(3), 04019111. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001763

[13] Qian, K., & Yahara, T. (2020). Mentality and behavior in COVID-19 emergency status in Japan: Influence of personality, morality and ideology. PloS one, 15(7), e0235883. https://doi.org/10.1371/journal.pone.0235883

[14] Al Ayub Ahmed, A., Aljabouh, A., Donepudi, P. K., & Suh Choi, M. (2021). Detecting Fake News Using Machine Learning: A Systematic Literature Review. arXiv-2102. https://doi.org/10.48550/arXiv.2102.04458

[15] Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032. https://doi.org/10.1016/j.mlwa.2021.100032

[16] Faustini, P. H. A., & Covoes, T. F. (2020). Fake news detection in multiple platforms and languages. Expert Systems with Applications, 158, 113503.

https://doi.org/10.1016/j.eswa.2020.113503

[17] Farokhian, M., Rafe, V., & Veisi, H. (2023). Fake news detection using dual BERT deep neural networks. Multimedia Tools and Applications, 1-18. https://doi.org/10.1007/s11042-023-17115-w

[18] Ibrishimova, M. D., & Li, K. F. (2020). A machine learning approach to fake news detection using knowledge verification and natural language processing. In Advances in Intelligent Networking and Collaborative Systems: The 11th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2019) (pp. 223-234). Springer International Publishing. https://doi.org/10.1007/978-3-030-29035-1_22

[19] Kula, S., Choraś, M., & Kozik, R. (2021). Application of the BERT-based architecture in fake news detection. In 13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12 (pp. 239-249). Springer International Publishing. https://doi.org/10.1007/978-3-030-57805-3_23

[20] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in neural information processing systems, 32. https://shorturl.at/mBHL7

[21] Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics, 46(2), 499-510. https://doi.org/10.1162/coli_a_00380

[22] Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. PeerJ Computer Science, 7, e443. https://doi.org/10.7717/peerj-cs.443

[23] Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018, August). Multi-source multi-class fake news detection. In Proceedings of the 27th international conference on computational linguistics (pp. 1546-1557).

[24] Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in cognitive sciences, 25(5), 388-402. https://doi.org/10.1016/j.tics.2021.02.007

[25] Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. International Journal of Information Management Data Insights, 1(2), 100052. https://doi.org/10.1016/j.jjimei.2021.100052

[26] Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. American Political Science Review, 115(3), 999-1015. https://doi.org/10.1017/S0003055421000290

[27] Igwebuike, E. E., & Chimuanya, L. (2021). Legitimating falsehood in social media: A discourse analysis of political fake news. Discourse & Communication, 15(1), 42-58. https://doi.org/10.1177/1750481320961659

[28] Li, Q., Hu, Q., Lu, Y., Yang, Y., & Cheng, J. (2020). Multi-level word features based on CNN for fake news detection in cultural communication. Personal and Ubiquitous Computing, 24, 259-272. https://doi.org/10.1007/s00779-019-01289-y

[29] Dabbous, A., Aoun Barakat, K., & de Quero Navarro, B. (2022). Fake news detection and social media trust: a cross-cultural perspective. Behaviour & Information Technology, 41(14), 2953-2972. https://doi.org/10.1080/0144929X.2021.1963475

[30] Petratos, P. N. (2021). Misinformation, disinformation, and fake news: Cyber risks to business. Business Horizons, 64(6), 763-774. https://doi.org/10.1016/j.bushor.2021.07.012

[31] Fong, B. (2021). Analysing the behavioural finance impact of'fake news' phenomena on financial markets: a representative agent model and empirical validation. Financial Innovation, 7(1), 1-30. https://doi.org/10.1186/s40854-021-00271-z

[32] Obadă, D. R., & Dabija, D. C. (2022). Can Fake News About Companies Lead to an Increased Social Media Usage? An Empirical Investigation. https://philpapers.org/rec/OBACFN

[33] Radwan, E., Radwan, A., & Radwan, W. (2020). The role of social media in spreading panic among primary and secondary school students during the COVID-19 pandemic: An online questionnaire study from the Gaza Strip, Palestine. Heliyon, 6(12). https://doi.org/10.1016/j.heliyon.2020.e05807

[34] Weiss, A. P., Alwan, A., Garcia, E. P., & Garcia, J. (2020). Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking. International Journal for Educational Integrity, 16, 1-30. https://doi.org/10.1007/s40979-019-0049-x

[35] Baptista, J. P., Correia, E., Gradim, A., & Piñeiro-Naval, V. (2021). The influence of political ideology on fake news belief: The Portuguese case. Publications, 9(2), 23. https://doi.org/10.3390/publications9020023

[36] Zanatta, E. T., Wanderley, G. P. D. M., Branco, I. K., Pereira, D., Kato, L. H., & Maluf, E. M. C. P. (2021). Fake news: The impact of the internet on population health. Revista da Associação Médica Brasileira, 67, 926-930. https://doi.org/10.1590/1806-9282.20201151

[37] Pérez-Escoda, A., Pedrero-Esteban, L. M., Rubio-Romero, J., & Jiménez-Narros, C. (2021). Fake news reaching young people on social networks: Distrust challenging media literacy. Publications, 9(2), 24. https://doi.org/10.3390/publications9020024

[38] Sampat, B., & Raj, S. (2022). Fake or real news? Understanding the gratifications and personality traits of individuals sharing fake news on social media platforms. Aslib Journal of Information Management, 74(5), 840-876. https://doi.org/10.1108/AJIM-08-2021-0232

[39] Torabi, M., & Sotudeh, H. (2022). The Role of Risk Perception and Ability to Detect Fake News in Acceptance of COVID-19 Vaccine among Students of Shiraz University, Iran. Health Information Management, 18(6), 265-271. https://doi.org/10.22122/him.v18i1.4440

[40] Chuai, Y., & Zhao, J. (2022). Anger can make fake news viral online. Frontiers in Physics, 10, 970174. https://doi.org/10.3389/fphy.2022.970174

[41] Tan, W. K., & Hsu, C. Y. (2023). The application of emotions, sharing motivations, and psychological distance in examining the intention to share COVID-19-related fake news. Online Information Review, 47(1), 59-80. https://doi.org/10.1108/OIR-08-2021-0448

[42] Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10, 178-206. https://doi.org/10.1162/tacl_a_00454

[43] Ghadiri, Z., Ranjbar, M., Ghanbarnejad, F., & Raeisi, S. (2022). Automated Fake News Detection using cross-checking with reliable sources. arXiv:2201.00083. https://doi.org/10.48550/arXiv.2201.00083

[44] Himdi, H., Weir, G., Assiri, F., & Al-Barhamtoshy, H. (2022). Arabic fake news detection based on textual analysis. Arabian Journal for Science and Engineering, 47(8), 10453-10469. https://doi.org/10.1007/s13369-021-06449-y

[45] Mazari, A. C., & Djeffal, A. (2022). Sentiment analysis of Algerian dialect using machine learning and deep learning with Word2vec. Informatica, 46(6). https://doi.org/10.31449/inf.v46i6.3340

[46] Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. AI Open. https://doi.org/10.1016/j.aiopen.2022.09.001

[47] Sallami, D., Gueddiche, A., & Aïmeur, E. (2023). From Hype to Reality: Revealing the Accuracy and Robustness of Transformer-Based Models for Fake News Detection. https://ceur-ws.org/Vol-3593/paper2.pdf

[48] de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., & Mattos, D. M. (2021). Identifying fake news on social networks based on natural language processing: trends and challenges. Information, 12(1), 38. https://doi.org/10.3390/info12010038

[49] Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. Disinformation, misinformation, and fake news in social media: Emerging research challenges and opportunities, 1-19. https://doi.org/10.1007/978-3-030-42699-6_1

[50] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

[51] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 30. https://doi.org/10.1007/s13278-023-01028-5

[52] Natarajan, R., Mehbodniya, A., Rane, K. P., Jindal, S., Hasan, M. F., Vives, L., & Bhatt, A. (2022). Intelligent gravitational search random forest algorithm for fake news detection. International Journal of Modern Physics C, 33(06), 2250084. https://doi.org/10.1142/S012918312250084X

[53] Hussain, M. G., Hasan, M. R., Rahman, M., Protim, J., & Al Hasan, S. (2020). Detection of bangla fake news using mnb and svm classifier. In 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 81-85). IEEE. https://doi.org/10.1109/iCCECE49321.2020.9231167

[54] Jain, P., Sharma, S., & Aggarwal, P. K. (2022). Classifying fake news detection using SVM, Naive Bayes and LSTM. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 460-464). IEEE. https://doi.org/10.1109/Confluence52989.2022.9734129

[55] Tourille, J., Sow, B., & Popescu, A. (2022, June). Automatic Detection of Bot-generated Tweets. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (pp. 44-51). https://doi.org/10.1145/3512732.3533584

[56] Aladeen, H. (2023). Breaking News: Machine Learning Helps to Spot Fake News Before it Spreads. IConFN'23, June 03–05, 2023, Wadiya.

[57] Biradar, S., Saumya, S., & Chauhan, A. (2023). Combating the infodemic: COVID-19 induced fake news recognition in social media networks. Complex & Intelligent Systems, 9(3), 2879-2891. https://doi.org/10.1007/s40747-022-00672-2

[58] Hutama, L. B., & Suhartono, D. (2022). Indonesian hoax news classification with multilingual transformer model and BERTopic. Informatica, 46(8). https://doi.org/10.31449/inf.v46i8.4336

[59] Adoma, A. F., Henry, N. M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 117-121). IEEE. https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379

[60] Wagh, V., Khandve, S., Joshi, I., Wani, A., Kale, G., & Joshi, R. (2021). Comparative study of long document classification. In TENCON 2021-2021 IEEE Region 10 Conference (TENCON) (pp. 732-737). IEEE. https://doi.org/10.1109/TENCON54134.2021.9707465

[61] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1), 100007. https://doi.org/10.1016/j.jjimei.2020.100007

[62] Yang, Z., Ma, J., Chen, H., Lin, H., Luo, Z., & Chang, Y. (2022). A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. arXiv:2209.14642. https://doi.org/10.48550/arXiv.2209.14642

[63] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M. Y., & Wang, W. Y. (2023). On the Risk of Misinformation Pollution with Large Language Models. arXiv:2305.13661. https://doi.org/10.48550/arXiv.2305.13661

[64] Sitaula, N., Mohan, C. K., Grygiel, J., Zhou, X., & Zafarani, R. (2020). Credibility-based fake news detection. Disinformation, misinformation, and fake news in social media: Emerging research challenges and Opportunities, 163-182. https://doi.org/10.1007/978-3-030-42699-6_9

[65] Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. Political Analysis, 29(1), 19-42. https://doi.org/10.1017/pan.2020.8

[66] Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165. https://shorturl.at/BJQX4

[67] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

[68] Yanagi, Y., Orihara, R., Sei, Y., Tahara, Y., & Ohsuga, A. (2020, July). Fake news detection with generated comments for news articles. In 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES) (pp. 85-90). IEEE. https://doi.org/10.1109/INES49302.2020.9147195

[69] Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. Journal of experimental political science, 9(1), 104-117. https://doi.org/10.1017/XPS.2020.37

[70] Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9291-9298). https://doi.org/10.1145/3581783.3612704

[71] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Wang, J. (2019). Release strategies and the social impacts of language models. arXiv:1908.09203. https://doi.org/10.48550/arXiv.1908.09203

[72] Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). Stylometric detection of ai-generated text in twitter timelines. arXiv:2303.03697. https://doi.org/10.48550/arXiv.2303.03697

[73] Najee-Ullah, A., Landeros, L., Balytskyi, Y., & Chang, S. Y. (2021). Towards detection of AI-generated texts and misinformation. In International Workshop on Socio-Technical Aspects in Security (pp. 194-205). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-10183-0_10

[74] Chang, S. Y. (2022). Towards detection of AI-generated texts and misinformation. In Socio-Technical Aspects in Security: 11th International Workshop, STAST 2021, Virtual Event (p. 194). Springer Nature. https://doi.org/10.1007/978-3-031-10183-0

[75] Gillioz, A., Casas, J., Mugellini, E., & Abou Khaled, O. (2020, September). Overview of the Transformer-based Models for NLP Tasks. In 2020 15th Conference on Computer Science and Information Systems (FedCSIS) (pp. 179-183). IEEE. https://doi.org/10.15439/2020F20

[76] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. https://shorturl.at/cCL23

[77] Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., ... & Miao, C. (2021). Diverse image inpainting with bidirectional and autoregressive transformers. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 69-78). https://doi.org/10.1145/3474085.3475436

[78] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2). https://shorturl.at/gyGMS

[79] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

[80] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

[81] Papapicco, C., Lamanna, I., & D'Errico, F. (2022). Adolescents' vulnerability to fake news and to racial hoaxes: a qualitative analysis on italian sample. Multimodal Technologies and Interaction, 6(3), 20. https://doi.org/10.3390/mti6030020

[82] Arin, K. P., Mazrekaj, D., & Thum, M. (2023). Ability of detecting and willingness to share fake news. Scientific Reports, 13(1), 7298. https://doi.org/10.1038/s41598-023-34402-6

[83] Peng, Y., Pei, C., Zheng, Y., Wang, J., Zhang, K., Zheng, Z., & Zhu, P. (2020). A cross-sectional survey of knowledge, attitude and practice associated with COVID-19 among undergraduate students in China. BMC public health, 20(1), 1-8. https://doi.org/10.1186/s12889-020-09392-z

[84] Mehta, Y., Majumder, N., Gelbukh, A., & Cambria, E. (2020). Recent trends in deep learning based personality detection. Artificial Intelligence Review, 53, 2313-2339. https://doi.org/10.1007/s10462-019-09770-z

[85] Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. Educational and psychological measurement, 80(3), 421-445. https://doi.org/10.1177/0013164419885

[86] Schreiber, J. B. (2022). Key processes and popular analyses in the SEM family of techniques. Contemporary Research Methods in Pharmacy and Health Services, 601-616. https://doi.org/10.1016/B978-0-323-91888-6.00023-5

[87] Zyphur, M. J., Bonner, C. V., & Tay, L. (2023). Structural equation modeling in organizational research: The state of our science and some proposals for its future. Annual Review of Organizational Psychology and Organizational Behavior, 10, 495-517. https://doi.org/10.1146/annurev-orgpsych-041621-031401

[88] Petropoulos, P., & Petropoulos, V. (2024). RoBERTa and Bi-LSTM for Human vs AI Generated Text Detection. Working Notes of CLEF.

[89] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. https://doi.org/10.48550/arXiv.1910.13461

[90] Kang, H. L., Na, S. D., & Kim, M. N. (2021). A method for enhancing speech and warning signals based on parallel convolutional neural networks in a noisy environment. Technology and Health Care, 29(S1), 141-152. https://doi.org/10.3233/THC-218015

[91] Cai, S., Han, D., Yin, X., Li, D., & Chang, C. C. (2022). A hybrid parallel deep learning model for efficient intrusion detection based on metric learning. Connection Science, 34(1), 551-577. https://doi.org/10.1080/09540091.2021.2024509

[92] Song, Y. F., Zhang, Z., Shan, C., & Wang, L. (2020, October). Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In proceedings of the 28th ACM international conference on multimedia (pp. 1625-1633). https://doi.org/10.1145/3394171.3413802

[93] Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. Journal of imaging, 6(6), 52. https://doi.org/10.3390/jimaging6060052

[94] Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research, 73, 329-396. https://doi.org/10.1613/jair.1.13200

[95] Singh, N., Kaliyar, R. K., Vivekanand, T., Uthkarsh, K., Mishra, V., & Goswami, A. (2022). B-LIAR: A novel model for handling Multiclass Fake News data utilizing a Transformer Encoder Stack-based architecture. In 2022 1st International Conference on Informatics (ICI) (pp. 31-35). IEEE. https://doi.org/10.1109/ICI53355.2022.9786925

[96] Nield, T. (2022). Essential Math for Data Science. " O'Reilly Media, Inc.". https://shorturl.at/ekvxH

[97] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv:1904.09675. https://doi.org/10.48550/arXiv.1904.09675

[98] Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. arXiv:2004.04696. https://doi.org/10.48550/arXiv.2004.04696

[99] Zhou, M., Li, Z., Tan, B., Zeng, G., Yang, W., He, X., ... & Xie, P. (2021). On the generation of medical dialogs for COVID-19. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). https://par.nsf.gov/servlets/purl/10345461

[100] Yuan, W., Neubig, G., & Liu, P. (2021). Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34, 27263-27277. https://shorturl.at/adqX0

[101] He, T., Zhang, J., Wang, T., Kumar, S., Cho, K., Glass, J., & Tsvetkov, Y. (2022). On the blind spots of model-based evaluation metrics for text generation. arXiv:2212.10020. https://doi.org/10.48550/arXiv.2212.10020

[102] Hunt, J. (2019). A beginners guide to Python 3 programming. Springer. https://doi.org/10.1007/978-3-030-20290-39

[103] Petrelli, M. (2021). Introduction to Python in Earth Science Data Analysis: From Descriptive Statistics to Machine Learning. Springer Nature. https://doi.org/10.1007/978-3-030-78055-5

[104] Bhattarai, B., Granmo, O. C., & Jiao, L. (2021). Explainable tsetlin machine framework for fake news detection with credibility score assessment. arXiv preprint arXiv:2105.09114. https://doi.org/10.48550/arXiv.2105.09114

[105] Guo, M., Liu, L., Guo, M., Liu, S., & Xu, Z. (2023). Accurate Generated Text Detection Based on Deep Layer-wise Relevance Propagation. In 2023 IEEE 8th International Conference on Big Data Analytics (ICBDA) (pp. 215-223). IEEE. https://doi.org/10.1109/ICBDA57405.2023.10104941