Generative AI for Explainable Automated Fact Checking on the FactEx: a New Benchmark Dataset

Saud Althabiti[1,2][0000-0002-4646-0577], Mohammad Ammar Alsalka[1][0000-0003-3335-1918], and Eric Atwell[1][0000-0001-9395-3764]

[1] University of Leeds, Leeds LS2 9JT, United Kingdom
[2] King Abdulaziz University, Jeddah 22254, Saudi Arabia
[1] scssal@leeds.ac.uk, [2] salthabiti@kau.edu.sa

Abstract. The immense volume of online information has made verifying claims' credibility more complex, increasing interest in automatic fact-checking models that classify evidence into binary or multi-class verdicts. However, there are few studies on predicting textual verdicts to explain claims' credibility. This field focuses on generating a textual verdict to explain a given claim based on a given news article. This paper presents our three-fold contribution to this field. Firstly, we collected the FactEx, an English dataset of facts with explanations from various fact-checking websites on different topics. Secondly, we employed seq2seq models and LLMs (namely T5, BERT2BERT, and BLOOM) to develop an automated fact-checking system. Lastly, we used ChatGPT to generate verdicts to check its performance and compared the results against other models. In addition, we explored the impact of dataset size on the model performance by conducting a series of experiments on seven different dataset sizes. The findings indicate that our fine-tuned T5-based model outperformed other generative LLMs and Seq2Seq Models with a ROUGE-1 score of about 26.75, making it the selected baseline for this task. Our study recommends examining the semantic similarity of the generative models for automatic fact-checking applications while also highlighting the importance of evaluating such models using additional techniques, such as crowed-based tools, to ensure the accuracy and reliability of the generated verdicts.

Keywords: FactEx Dataset , Automatic Fact-check, ChatGPT, Generative LLMs , NLP, Artificial Intelligence, Computer Science, Disinformation

1 Introduction

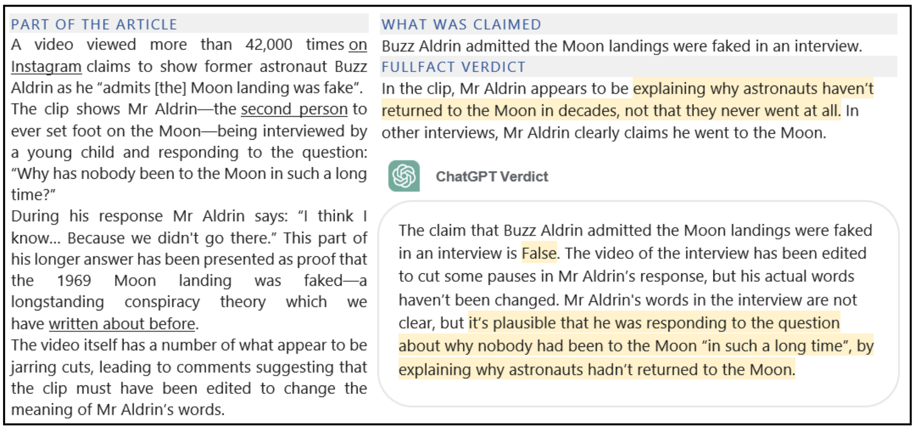
Fake news is a form of false information that can be intentionally spread through various media sources, such as traditional media, social media, or news websites [1]. The purpose of fake news is often to manipulate public opinion or beliefs, typically for political or financial gain [2]–[4]. It can be spread by individuals, organizations, or even governments to discredit opponents or promote their interests [4]–[6]. The consequences of fake news can be serious, creating confusion and mistrust, causing discord among different groups, and even inciting violence [7], [8]. Hence, it is essential to remain vigilant and critical of the information we consume, particularly online, and fact-check sources to differentiate between real and fake news.

This is where fact-checking websites play a crucial role in ensuring the information presented to the public is accurate and reliable. Fact-checking websites provide a platform for individuals to verify the authenticity of news and information by checking sources and validating claims [9]. These websites not only help in maintaining the integrity of information but also in educating people about how to identify fake news and misinformation. With the rise of social media and the increasing spread of fake news, the need for fact-checking websites has become more necessary than ever before [10]. One of the most popular fact-checking organizations is PolitiFact.com. It offers a rating system that assesses the accuracy of factual claims, including True, Half True, False, and "Pants on Fire" [11]. Another valuable way of fact-checking involves investigators examining related data and documents to evaluate claims and then disseminate their verdicts to the public, such as Fullfact.org.

However, manual fact-checking is tedious and too slow to keep up with the speed of online information. To address this, the journalism community can benefit from automating the fact-checking process using AI and NLP tools. This will help validate large amounts of new details that appear and spread quickly, motivating the need for automated fact-checking systems [10].

Although many systems have focused on binary or multi-class classification problems, such as predicting a binary verdict from text [10], [12], [13], this study investigates the use of sequence-to-sequence (Seq2Seq) and large language models (LLMs) to predict a textual verdict that explains a given claim based on a given article.

In our objective of generating explanations, we were initially motivated to explore one of the trending LLMs, such as ChatGPT, to see if it could justify a claim from the information provided in an article. As shown in Fig. 1, the generated verdict closely aligns with the human-written explanation on the trusted FullFact.org website. This shows the potential of such models to deliver reliable and comparable explanations, supporting our goal of advancing the field of explainable automatic fact-checking.



PART OF THE ARTICLE

A video viewed more than 42,000 times on Instagram claims to show former astronaut Buzz Aldrin as he "admits [the] Moon landing was fake". The clip shows Mr Aldrin—the second person to ever set foot on the Moon—being interviewed by a young child and responding to the question: "Why has nobody been to the Moon in such a long time?"

During his response Mr Aldrin says: "I think I know… Because we didn't go there." This part of his longer answer has been presented as proof that the 1969 Moon landing was faked—a longstanding conspiracy theory which we have written about before.

The video itself has a number of what appear to be jarring cuts, leading to comments suggesting that the clip must have been edited to change the meaning of Mr Aldrin's words.

WHAT WAS CLAIMED

Buzz Aldrin admitted the Moon landings were faked in an interview.

FULLFACT VERDICT

In the clip, Mr Aldrin appears to be explaining why astronauts haven't returned to the Moon in decades, not that they never went at all. In other interviews, Mr Aldrin clearly claims he went to the Moon.

ChatGPT Verdict

The claim that Buzz Aldrin admitted the Moon landings were faked in an interview is False. The video of the interview has been edited to cut some pauses in Mr Aldrin's response, but his actual words haven't been changed. Mr Aldrin's words in the interview are not clear, but it's plausible that he was responding to the question about why nobody had been to the Moon "in such a long time", by explaining why astronauts hadn't returned to the Moon.

Fig. 1. Example from FullFact.org and a generated verdict using ChatGPT.

The purpose of this paper is to present our three-fold contribution, which includes:

- Collecting FactEx, a new English dataset for fact-checking explanations from trusted websites containing news articles, claims, and corresponding textual verdicts.
- We secondly fine-tuned some LLMs and seq2seq models, namely T5, BLOOM, and BERT2BERT architecture, to develop an automatic explainable fact-checking system and compare the results obtained from these models. The best-performing one is then subsequently published. To the best of our knowledge, we are the first to explore such effective architectures for this purpose.
- Last but not least, we attempt to consider a sample of our dataset to evaluate ChatGPT's capabilities by generating verdicts. We compare the results with other models to measure performances using the ROUGE scores.

The subsequent section includes a literature review of fake news detection and related work. Section three describes the methodology, including dataset collection, preprocessing, applying seq2seq models, and the evaluation method used. The results and discussion of the conducted experiments are detailed in the fourth section. Finally, we conclude this paper and suggest future work.

2 Related work

Many NLP studies commonly view claim verification as a text classification task by building models that analyze a claim under investigation along with its retrieved evidence in order to reach a verdict regarding the claim. This verdict can typically be classified into different categories, such as support, contradict, or not enough information [14]–[19]. To implement classification tasks, various methods were used, including traditional machine learning algorithms, deep learning models, and Transformer-based models. These methods typically involve feature engineering and modeling steps, where text data is pre-processed, features are extracted, and a classification model is trained on labeled data [20], [21]. While in our study, we focus on seq2seq pre-trained models to provide an explanation rather than just a specific category.

Since the presence of textual justifications from journalists to explain verdicts is scarce in most available datasets [10], the study [22] expanded the LIAR dataset [11] by incorporating human justifications extracted from fact-checking articles. Although these justifications were initially intended as additional information to support claim verification and improve both binary and multi-classification tasks, it was also used by [23] to generate summaries. They employ an extractive method to generate justification summaries using DistilBERT. In contrast, the paper [24] adopted a joint approach involving both extractive and abstractive summarization. Additionally, they introduced the first dataset, which includes explanations crafted by journalists, fact-checking articles, and other news items related to public health claims [25]. Furthermore, [26] used the FEVER dataset [27] and a GPT-3-based system to generate summaries, resulting in a new dataset called e-FEVER consisting of 67,687 examples. On the other hand, this is the first study that investigates Seq2Seq models and compares them with generative large language models, such as ChatGPT, to generate claim verification explanations.

Table 1 summarizes these studies regarding the utilized datasets and employed methods.

Table 1. Comparative of Related Studies: Datasets, Explanations (Ex), and Methodologies

Study	Size	Topics	Explained by	Ex. Source	Model
[22]	12,836	Various	Humans	PolitiFact	ML models
[23]	12,836	Various	Humans	PolitiFact	DistilBERT
[26]	67’687	Various	Generated	Generated	GPT-3
[25]	11,832	Health	Humans	Various	BERT
FactEx	12,150	Various	Humans	Various	T5, BERT2BERT, BLOOM, and ChatGPT

3 Methodology

3.1 The FactEx Dataset

In order to train Seq2Seq models to predict explanations for fact-checking verdicts, we needed a dataset that combines claims, articles, and corresponding judgments. Therefore, we collected a new dataset named “**FactEx**”[1] (**Fact Ex**plained). The dataset contains 12,150 records from three trusted fact-checking websites, namely, FullFact.org, PolitiFact.com, and BBC.co.uk, spanned from 2016 and 2023. This ensures that our dataset contains the most recent and relevant information from various reliable sources on different topics, such as health, economy, politics, education, and more. Initially, we used Google's fact-checking tool API, a tool that allows us to search for fact-checks previously published by fact-checking organizations, which provides a structured JASN file, as shown in Fig. 2. This streamlines the dataset collection process by handling JSON formatting.

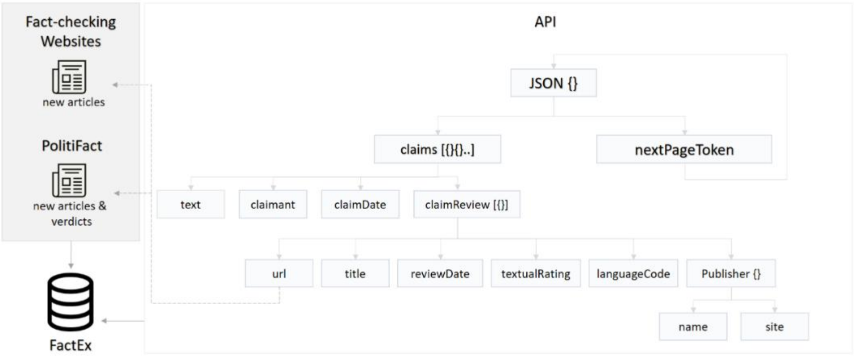


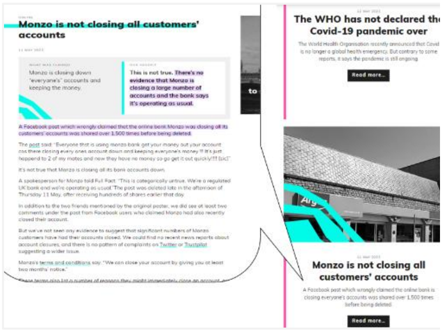
Fig. 2. The FactEx collection process using google API and NLP tools.

[1] https://github.com/althabiti/FactEx

We also collected the full articles related to each claim to make our dataset more informative by using the provided URLs with each claim, which yielded the following features to our dataset:

- URL [*string*]: The URL associated with the article.
- Title [*string*]: The title of the article.
- Text [*string*]: the claim text.
- TextualRating[*string*]: The verdict.
- Article [*string*]: The text content of the article.
- Article_HTML [*string*]: The text content of the article, including the HTML tags.
- Additional features such as, claimDate, claimant, and reviewDate.

The FactEx dataset becomes more diverse and reliable by including content from different trustworthy sources. This makes it a valuable resource for researchers and practitioners working on automated fact-checking systems. Fig. 3 provides an example from the FullFact.org website showcasing a claim, the related article, and a journalist's explanation of the verdict. While Fig. 4 presents a PolitiFact claim example and the relative website structure, which includes the title, article, and the verdict explanation, starting from the "Our ruling" section.



Monzo is not closing all customers' accounts

11 MAY 2023

Monzo is closing down "everyone's" accounts and keeping the money.

This is not true. There's no evidence that Monzo is closing a large number of accounts and the bank says it's operating as usual.

A Facebook post which wrongly claimed that the online bank Monzo was closing all its customers' accounts was shared over 1,500 times before being deleted.

The post said: "Everyone that is using monzo bank get your money out your account cos there closing every ones account down and keeping everyone's money !! It's just happend to 2 of my mates and now they have no money so go get it out quickly!!! [sic]"

It's not true that Monzo is closing all its bank accounts down.

A spokesperson for Monzo told Full Fact: "This is categorically untrue. We're a regulated UK bank and we're operating as usual." The post was deleted late in the afternoon of Thursday 11 May, after receiving hundreds of shares earlier that day.

In addition to the two friends mentioned by the original poster, we did see at least two comments under the post from Facebook users who claimed Monzo had also recently closed their account.

But we've not seen any evidence to suggest that significant numbers of Monzo customers have had their accounts closed. We could find no recent news reports about account closures, and there is no pattern of complaints on Twitter or Trustpilot suggesting a wider issue.

Monzo's terms and conditions say: "We can close your account by giving you at least two months' notice."

These terms also list a number of reasons they might immediately close an account...

The WHO has not declared th

Covid-19 pandemic over

The World Health Organisation recently announced that Covid is no longer a global health emergency. But contrary to some reports, it says the pandemic is still ongoing.

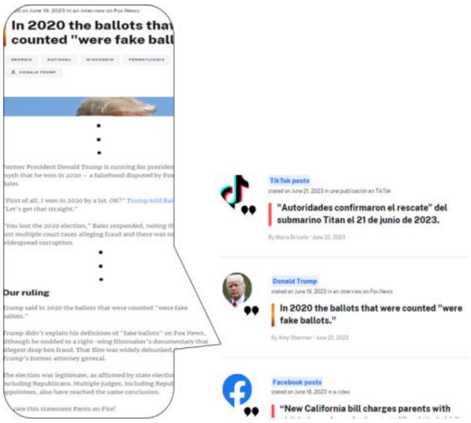
Read more...

11 MAY 2023

Monzo is not closing all customers' accounts

A Facebook post which wrongly claimed the online bank is closing everyone's accounts was shared over 1,500 times before being deleted.

Read more...



d on June 19, 2023 in an interview on Fox News:

In 2020 the ballots tha counted "were fake ball

GEORGIA NATIONAL WISCONSIN PENNSYLVANIA

DONALD TRUMP

Former President Donald Trump is running for presiden myth that he won in 2020 — a falsehood disputed by Fo Baier.

"First of all, I won in 2020 by a lot. OK?" Trump told Ba "Let's get that straight."

"You lost the 2020 election," Baier responded, noting th lost multiple court cases alleging fraud and there was n widespread corruption.

Our ruling

Trump said in 2020 the ballots that were counted "were fake ballots."

Trump didn't explain his definition of "fake ballots" on Fox News, although he nodded to a right-wing filmmaker's documentary that alleged drop box fraud. That film was widely debunked, Trump's former attorney general.

The election was legitimate, as affirmed by state electio including Republicans. Multiple judges, including Repu appointees, also have reached the same conclusion.

ate this statement Pants on Fire!

TikTok posts

stated on June 21, 2023 in una publicación en TikTok

"Autoridades confirmaron el rescate" del submarino Titan el 21 de junio de 2023.

By Maria Briceño · June 22, 2023

Donald Trump

stated on June 19, 2023 in an interview on Fox News

In 2020 the ballots that were counted "were fake ballots."

By Amy Sherman · June 22, 2023

Facebook posts

stated on June 18, 2023 in a video

"New California bill charges parents with

Fig. 3. FullFact.org example

Fig. 4. PlitiFact.com example

While retrieving the data, we encountered a challenge related to the inconsistent structure of the web pages. More specifically, the PolitiFact website has different styles that exhibit variations in the HTML tags and classes used to present articles and verdicts. This presented a significant obstacle in accurately extracting the desired content. To overcome this challenge, we adopted a two-step approach. Initially, we employed the BeautifulSoup library to scrape the entire web page, encompassing all HTML tags and content. Subsequently, we utilized NLP tools to selectively extract the relevant information, such as articles and verdicts, while filtering out irrelevant elements. This process allowed us to facilitate the impact of varying webpage structures and ensured the inclusion of all necessary information for our dataset. Furthermore, we included the

URLs and HTML files in the dataset, enabling future enhancements as we aim to contribute to advancing research in the field of automated fact-checking. In addition, we excluded instances where the web pages contained lengthy explanations without explicitly mentioning the verdicts.

3.2 Preprocessing and Methods

In this study, we initially experimented with 900 claims samples from the collected FactEx dataset. We first split our dataset into three sets. Training: to train the model parameters; validation: for tuning hyperparameters; and testing: to check the performance of the tuned model. The sizes of the split data are 600, 150, and 150, respectively. Additional texts were prepended to each sample to help the selected models distinguish the contextual cues and establish a clear pattern for all samples. For example, *"claim:"* was prepended before each claim and *"article:"* before each new article. Secondly, a common practice when applying transformer models is to tokenize inputs and outputs. In this case, the article and a claim are the input, and the verdict is the target source we aim to predict.

It is generally challenging to train transformer-based models from scratch, requiring extensive datasets and high GPU memory. Therefore, to conduct the study, we decided to investigate four different models including T5, BERT2BERT, Bloom, and ChatGPT.

T5 (**T**ext-**T**o-**T**ext **T**ransfer **T**ransformer) is a transformer-based language model developed by Google AI Language [28]. It is pre-trained on various natural language tasks using a text-to-text format, where the input and output are both text strings. As a result, T5 has achieved state-of-the-art results on various natural language processing tasks such as question answering, text summarization, and language translation [28]. We fine-tuned the T5-base model with learning rates of 4e-5 and 3 epochs.

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is a powerful pre-trained encoder model that can be used to create a fix-sized representation of the text [29]. To use the model as a decoder, we followed the steps in a demonstrated architecture [30] that uses BERT to create an encoder-decoder architecture (BERT2BERT) for seq2seq models. We then fine-tuned the presented architecture on the 900 samples dataset using the default hyperparameters.

BLOOM is another open-source alternative for text generation. It is a recently released transformer-based large language model with about 176 billion parameters. We evaluated its performance on the 900 samples and compared it with the Seq2Seq models. Fig. 5 illustrates the process structure of our experiments.

GPT-3 (**G**enerative **P**re-trained **T**ransformer **3**) is a state-of-the-art language model developed by OpenAI [31]. It is a transformer-based language model that is trained on a massive corpus of diverse natural language data to generate human-like text.

ChatGPT is a fine-tuned model using reinforcement learning based on GPT-3 architecture [32]. It has a broad range of language capabilities, including language translation, question answering, text completion, and text summarization. OpenAI provides a full guide on how to fine-tune their models. Using their API, we integrated the "text-davinci-003" model and set the parameter "temperature" equal to 7 to increase the randomness of the generated texts [33], as we aim to predict an explanation rather than just a unique answer. Generative models, such as GPT3-based models can be effectively employed with minimal modifications, with or without fine-tuning as demonstrated by [26]. Therefore, we tested ChatGPT to see if it could provide a sound explanation when providing both the articles along with the claim on 180 samples with few-shot learning.

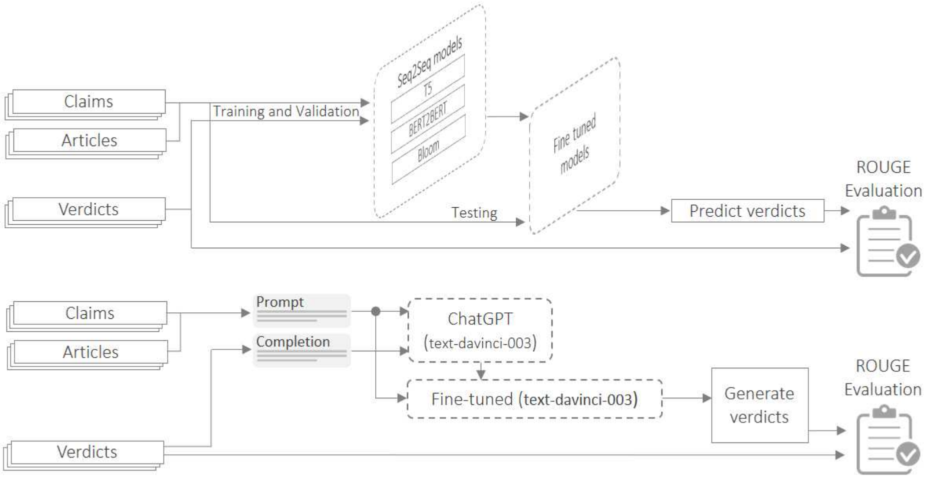


Fig. 5. Methodology architecture

To fine-tune the model, two main things should be provided: prompt and completion. Within each prompt *P*, we instructed the model to follow the steps that should be considered for the generation, along with claims *C* and articles *A* and provided the verdicts *V* to be the completion appended with an ending tag. After training 20 samples, we tested the fine-tuned ChatGPT model on 160 samples to generate explanations by providing prompts, as the explained pseudocode in Algorithm 1, including the instruction *I*, claims, and articles only.

Algorithm 1: Prompt used to instruct the fine-tuned generative model

```
Input: $C_i$, $A_i$ where $i \in FactEx$
Output: $V_i$

$I$ ← "Given a text article starting from 'text_article:' and a claim starting from 'claim:', suggest a verdict based on possible evidence retrieved from the article."
for $i = 1$ to $n$
    $M_i$ ← [{"role": "system", "content": "You are an automatic Fact Checker acting like a journalist" + $I$ + $A_i$ + $C_i$ }]
    $P_i$ ← ({"role": "assistant", "content": $M_i$})
    $V_i$ ← getCompletion($P_i$)
end for
```

3.3 Evaluation

The two widely used text generation tasks are machine translation and text summarization, evaluated by BLEU and ROUGE scores, respectively. BLEU (Bilingual Evaluation Understudy) is mainly used for evaluating machine translation systems. It calculates how well the generated translation aligns with one or more reference translations [34]. In contrast, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a more general metric for evaluating various NLP tasks, such as text summarization [35]. In our task, the verdict generation is comparable to text summarization as it aims to convey the article's essence to the reader; hence we will evaluate our results using the ROUGE score [35], [36].

4 Results and Discussion

4.1 LLMs and Seq2Seq Results Comparison

As we initially split the dataset to train and validate, we tested them on the fine-tuned models. The results of the predicted verdicts are evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores. ROUGE-1 is the overlapping of unigram or each word between the human verdicts and the predicted explanations, ROUGE-2 is the overlapping of bigrams, and ROUGE-L and ROUGE-Lsum are calculating the longest common subsequent to capture sentence structure. While the ROUGE-L is computed as the average of individual sentences, the ROUGE-Lsum is calculated over the whole predicted text [35].

One of the objectives of this paper is to test a sample of claims and articles to generate verdicts using ChatGPT to compare its performance with journalists' verdicts and other Seq2Seq methods. Table 2 indicates that the fine-tuned T5 model outperformed other models when evaluating using the ROUGE score.

Table 2. Testing results using ROUGE metrics.

Used Model	rouge1	rouge2	rougeL	rougeLsum
Our T5-based model	**26.75**	**10.45**	**21.95**	**23.43**
BERT2BERT	18.89	04.07	14.18	14.22
Bloom	03.54	01.84	02.84	03.24
ChatGPT	10.87	01.67	08.57	08.65

4.2 Model's Performance vs Dataset Size

To explore the impact of dataset size on the model performance, we specifically focused on investigating the T5-small. We conducted a series of experiments on different dataset sizes, ranging from 900 to 1500 samples. We fine-tuned using approximately four-sixths of the dataset for training, one-sixth for validation, and the remaining for testing in each case and calculated ROUGE metrics. Upon analysis, we observed that the results exhibited a slight fluctuated increase with no significant difference among the dataset sizes tested. ROUGE-1 scores, for instance, range from approximately 23.4 to 25.3, as shown in Fig. 6. Given this, increasing the dataset size does not consistently lead to a considerable improvement in the model's performance. Therefore, we decided to use our trained T5-based model, presented in Table 2 as a baseline model for the task of automatic textual fact-checking explanations. Our model can be found on the HuggingFace.co repository[2].

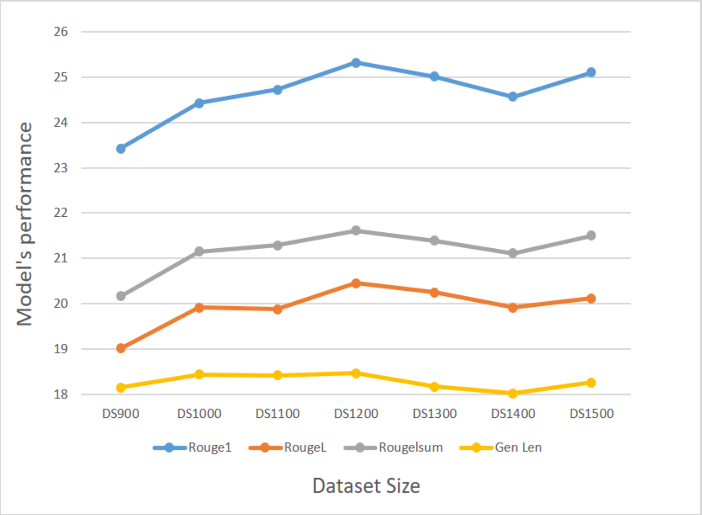


Fig. 6. The impact of dataset size on the model performance.

Table 3 presents two examples, each featuring a verdict and its source. In the first example, FullFact determined that the claim was true, and our model successfully classified the overall truthfulness of the claim, unlike the BERT2BERT model, which discredited the claim's credibility from the beginning. On the other hand, the second example compares a FullFact judgment with a verdict generated by ChatGPT. Although the claim is accurate for part of the claim, as the FullFact deemed, the generated text explained that by stating "Partly true", some of the chosen words align with those of the FullFact.

Despite the low scores, these instances showed promising outcomes. As seen in Table 3, T5 generates exact word matches, such as "correct" and "Switzerland", whereas the others generate meanings that may be more or less related but not the exact words. Since the ROUGE metrics are based on the same exact word matches and compute the overlap of n-grams (consecutive words of length n), it led to a significant difference between the results. Therefore, the meaning of the entire sentence must be considered in future evaluations.

[2] https://huggingface.co/althabiti/VerdictGen_t5-based

Table 3. Two examples of human verdicts compared with generated verdicts.

Source	Verdict
FullFact[3]	That's correct. Switzerland has some access to the EU's single market. It pays financially for this and takes on certain EU laws
Our T5-based model	It is correct. The EU imported a 20,000 of goods and services per person from Switzerland
BERT2BERT	It is not true, but it does not necessarily mean it would have to be used to contradict the uk. but it does not mean it can be used as a currency, …
FullFact[4]	This claim does not factor in people who identified as white but not white British, and so is not true for either London or Manchester. It is accurate for Birmingham, where 48.6% identified as white.
GPT-3 based model	Verdict: Partly true: Minority group identified surveyed based across empty ethnic general become more usual England constituent cities nation.

5 Conclusion and Future Work

As online information increases continuously, it has become increasingly challenging for individuals to verify the truthfulness of claims they encounter. To address this problem, there is a growing interest in developing automatic fact-checking models that can analyze textual evidence and classify them into binary verdicts about the veracity of claims, for example, "True or False". However, fewer studies explored the problem of predicting textual explanations of claim credibility.

This paper has three main contributions, as we aim to develop an explainable automatic fact-checking model to assess the truthfulness of claims based on supporting articles. To achieve this goal, we first created the FactEx, a new dataset containing 12,150 samples on different topics from three trusted fact-checking websites. Each sample has various features, including a claim and a verdict (an explanation) paired with a corresponding article to serve as evidence for our model.

We then applied a seq2seq architecture to generate explanations for each claim by fine-tuning our models to achieve better performance. In the process, we conducted a comparison of different generative LLMS and seq2seq models, namely, T5, BERT2BERT, BLOOM, and ChatGPT, by evaluating their ROUGE scores. Based on our findings, we observed that the fine-tuned T5-based model outperforms other models with about 26.75 ROUGE1 score and made it publicly available for future use as a baseline model for this task. On the other hand, the discussion recommends investigating the semantic similarities rather than just the syntactic for the generative models, such as ChatGPT, which have strong potential for use in automatic fact-checking applications. We also concluded that increasing the dataset size does not always lead to a considerable improvement in the model's performance, as we utilized the T5-small model across seven different dataset size attempts.

[3] https://fullfact.org/europe/vote-leave-facts-leaflet-exports/

[4] https://fullfact.org/immigration/nigel-farage-census-london-manchester/

While there is still much room for improvement in model robustness and evaluation technique, the results of this study provide a strong foundation for future research in this area. We also aim to extend this methodology to other languages, such as Arabic, since there are fewer fact-checking websites. In terms of evaluation, conducting a comprehensive human assessment to evaluate the extent to which ROUGE scores align with semantic similarity would be valuable. This could involve engaging experts in the field of misinformation, including journalists, social scientists, and politicians, to provide a more nuanced understanding of the quality of our model's explanations. As our main focus is on the automation parts of the task, joint efforts in this direction could significantly contribute to a deeper understanding of the relationship between automated metrics like ROUGE and human judgment, including assessing its accuracy and coherence. Furthermore, we aim to create a crowdsourcing tool for users to get a larger pool of evaluators to determine the generated verdicts and provide feedback.

References

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

[2] E. C. Tandoc Jr, Z. W. Lim, and R. Ling, "Defining 'fake news' A typology of scholarly definitions," *Digital journalism*, vol. 6, no. 2, pp. 137–153, 2018.

[3] M. R. Jahng, H. Lee, and A. Rochadiat, "Public relations practitioners' management of fake news: Exploring key elements and acts of information authentication," *Public Relat Rev*, vol. 46, no. 2, p. 101907, 2020.

[4] S. Tejedor, M. Portalés-Oliva, R. Carniel-Bugs, and L. Cervi, "Journalism students and information consumption in the era of fake news," *Media Commun*, vol. 9, no. 1, pp. 338–350, 2021.

[5] A. Andorfer, "Spreading like wildfire: Solutions for abating the fake news problem on social media via technology controls and government regulation," *Hastings LJ*, vol. 69, p. 1409, 2017.

[6] M. Rapti, G. Tsakalidis, S. Petridou, and K. Vergidis, "Fake News Incidents through the Lens of the DCAM Disinformation Blueprint," *Information*, vol. 13, no. 7, p. 306, 2022.

[7] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, "The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review," *J Public Health (Bangkok)*, pp. 1–10, 2021.

[8] S. A. Khan, M. H. Alkawaz, and H. M. Zangana, "The use and abuse of social media for spreading fake news," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, IEEE, 2019, pp. 145–148.

[9] D. Zlatkova, P. Nakov, and I. Koychev, "Fact-checking meets fauxtography: Verifying claims about images," *arXiv preprint arXiv:1908.11722*, 2019.

[10] Z. Guo, M. Schlichtkrull, and A. Vlachos, “A survey on automated fact-checking,” *Trans Assoc Comput Linguist*, vol. 10, pp. 178–206, 2022.

[11] W. Y. Wang, “‘ liar, liar pants on fire’: A new benchmark dataset for fake news detection,” *arXiv preprint arXiv:1705.00648*, 2017.

[12] X. Zeng, A. S. Abumansour, and A. Zubiaga, “Automated fact-checking: A survey,” *Lang Linguist Compass*, vol. 15, no. 10, p. e12438, 2021.

[13] N. Naderi and G. Hirst, “Automated fact-checking of claims in argumentative parliamentary debates,” in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 60–65.

[14] S. S. Alanazi and M. B. Khan, “Arabic Fake News Detection In Social Media Using Readers’ Comments: Text Mining Techniques In Action,” *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY*, 2020.

[15] S. Althabiti, M. Alsalka, and E. Atwell, “SCUoL at CheckThat! 2021: An AraBERT model for check-worthiness of Arabic tweets,” in *CEUR Workshop Proceedings*, 2021.

[16] J. K ̈ohler, M. Shahi, Gautam Kishore Struß, Julia Maria Wiegand, M. Siegel, T. Mandl, and M. Sch ̈utz, “Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection,” in *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*, Bologna, Italy, 2022.

[17] P. Nakov *et al.*, “Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection,” in *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, Bologna, Italy, 2022.

[18] P. Nakov *et al.*, “The CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection,” in *European Conference on Information Retrieval*, Springer, 2022, pp. 416–428.

[19] S. Althabiti, M. A. Alsalka, and E. Atwell, “SCUoL at CheckThat! 2022: fake news detection using transformer-based models,” in *CEUR Workshop Proceedings*, CEUR Workshop Proceedings, 2022, pp. 428–433.

[20] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” *ACM Comput Surv*, 2020, doi: 10.1145/3395046.

[21] S. Althabiti, M. A. Alsalka, and E. Atwell, “Detecting Arabic Fake News on Social Media using Sarcasm and Hate Speech in Comments”.

[22] T. Alhindi, S. Petridis, and S. Muresan, “Where is your evidence: Improving fact-checking by justification modeling,” in *Proceedings of the first workshop on fact extraction and verification (FEVER)*, 2018, pp. 85–90.

[23] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, “Generating fact checking explanations,” *arXiv preprint arXiv:2004.05773*, 2020.

[24] N. Kotonya and F. Toni, “Explainable automated fact-checking: A survey,” *arXiv preprint arXiv:2011.03870*, 2020.

[25] N. Kotonya and F. Toni, “Explainable automated fact-checking for public health claims,” *arXiv preprint arXiv:2010.09926*, 2020.

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

[27] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "Fever: a large-scale dataset for fact extraction and verification," *arXiv preprint arXiv:1803.05355*, 2018.

[28] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

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

[30] Ala Alam Falaki, "How To Train a Seq2Seq Summarization Model Using 'BERT' as Both Encoder and Decoder!! (BERT2BERT)," 2022. https://pub.towardsai.net/how-to-train-a-seq2seq-summarization-model-using-bert-as-both-encoder-and-decoder-bert2bert-2a5fb36559b8

[31] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach (Dordr)*, vol. 30, pp. 681–694, 2020.

[32] H. Hassani and E. S. Silva, "The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field," *Big data and cognitive computing*, vol. 7, no. 2, p. 62, 2023.

[33] M. Zong and B. Krishnamachari, "a survey on GPT-3," *arXiv preprint arXiv:2212.00857*, 2022.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[35] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[36] Y. M. Wazery, M. E. Saleh, A. Alharbi, and A. A. Ali, "Abstractive Arabic text summarization based on deep learning," *Comput Intell Neurosci*, vol. 2022, 2022.