TRUExT: Trustworthiness Regressor Unified Explainable Tool

Catherine Donner
Data Science and Analytics Institute
University of Oklahoma
Norman, Oklahoma
Catherine.G.Donner-1@ou.edu

Gopichandh Danala
Data Institute for Societal Challenges
University of Oklahoma
Norman, Oklahoma
danala@ou.edu

Wolfgang Jentner
Data Institute for Societal Challenges
University of Oklahoma
Norman, Oklahoma
wjentner@ou.edu

David Ebert
Data Institute for Societal Challenges
University of Oklahoma
Norman, Oklahoma
ebert@ou.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825100

***Abstract*—One of the most pressing public policy issues that has involved transdisciplinary research in the field of data science is rapidly detecting widespread misinformation. While data science can pose a lot of potential for solving the big-data problem of misinformation on an automated scale, it likewise requires insights from the field of communications and journalism to define quantifiable features that can assist in more accurate misinformation predictions. Currently, the preeminent tools used for misinformation detection are large language models (LLMs) as they are renowned for their ability to capture the context and meaning of textual data. However, despite advancements in developing effective data science models and tools for identifying misinformation, there are not many available options for evaluating news article content for misinformation potential. This study proposes TRUExT, an explainable, regression-based data tool that integrates multiple communication-based natural language processing (NLP) dimensions with a base LLM to holistically evaluate trustworthiness in news articles. It was found that the Hugging Face LLM RoBERTa with the added NLP dimensions as features was the most effective foundational model after testing multiple LLMs. Furthermore, TRUExT introduced a potential big-data solution to the growing problem of misinformation through research intersecting data science and communications to capture not only the technicality of misinformation data predictions but also certain communication factors in the data. In the future, this tool could likewise be deployed to be used by U.S.-based stakeholders who have an important role in the ongoing information war.**

***Index Terms*—misinformation, large language models (LLMs), natural language processing (NLP), communications and journalism**

I. Introduction

Misinformation is defined as an incorrect or misleading statement that can obscure the truth [16], while similarly disinformation is defined as false information deliberately created or spread in order to cause harm, usually with political, psychological, or social motivations [13]. Ideally, widespread media literacy and regulations on social media content are the most effective methods for helping stop the spread of misinformation and disinformation [9]. However, without these desired regulations currently in place, there is a great demand for handling massive amounts of misinformation through big-data analytical tools that can quickly and accurately classify misinformation. As a field, data science can have the technological ability to attain this difficult objective.

For background context, a previous project required the devising of a framework for creating trustworthiness scores based on certain dimensions of news article texts. For databases of information that needed to be populated with scores at an efficient pace, application programming interfaces (APIs) were examples of such industry-effective tools used to populate fields of information, as these can provide multi-faceted information about input entities in a quick response time. One API that was explored was the Romanian-based Zetta Cloud TrustServista API[1], which provided misinformation scores as well as explanations for what dimensions were considered for the scores (i.e., named entities, clickbait, sentiment). However, due to its lack of transparency on cost, this API was not considered in final implementations for the project. Besides the TrustServista API, very few APIs were available that could perform the intended task of wholesomely determining a trustworthiness score for any given text, regardless of length. For those options that were available, these were usually generative AI chatbot models like ChatGPT in API form. Thus, the lack of availability of efficient automated misinformation detection tools served as the fundamental motive of this study.

The objective of this study was to create and propose a data tool that integrated a large language model (LLM) as the tool foundation with communication-based dimensions converted into natural language processing (NLP) features as supporting explainability and prediction factors. This tool was named TRUExT, which stands for Trustworthiness Regressor Unified Explainable Tool. The tool used news article texts as inputs

Data Institute for Societal Challenges, University of Oklahoma

[1]https://www.trustservista.com/trustservista-api/

and cumulative regression-based trustworthiness scores on a scale from 0 to 100 (0 meaning not trustworthy and high misinformation potential and 100 meaning very trustworthy and low misinformation potential) as outputs. A compiled training dataset of news article texts and titles from a wide diversity of origins was used as the data for this study (no social media data from Facebook, X, etc. was used). Multiple open-source LLMs from the Hugging Face website[2] including BERT, RoBERTa, and DistilBERT [11], [20], [30], [41] were fine-tuned for the intended regression task. These LLMs were analyzed and compared to select the optimal foundation model for TRUExT. This study showed that the RoBERTa model significantly had the highest Pearson and Spearman Rank correlation coefficients as well as the lowest mean squared error (MSE) losses out of the 3 models tested, therefore making it the proposed base model for TRUExT. TRUExT with the RoBERTa model was also shown to perform better than baseline algorithmic predictions using NLP dimensions of misinformation.

TRUExT will help contribute a new, efficient data science tool to the current competition of very limited available tools specifically used for predicting trustworthiness (or misinformation potential) scores on a regression-based scale, for news article texts originating from a wide variety of sources. TRUExT incorporates multiple communication-based NLP dimensions as explainability measures, contributing to explainable artificial intelligence (XAI) and establishing trustworthiness in such tools for potential users.

Additionally, this study has potential societal impacts on several targeted stakeholders outlined below:

- Military agencies (i.e., U.S. Air Force): Consistent continuation of supply chain operations for the U.S. military by holding misinformation accountable that can sabotage supplier business decisions.
- Government agencies (i.e., DoD, DHS, CIA, NSA): U.S. government agencies can utilize effective data tools like TRUExT for automated misinformation detection in foreign and domestic news media and for tackling misinformation on a big-data scale in the ongoing information war. Broader implications can include protection of national security interests and protection of U.S. citizens and agencies from damaging misinformation.
- Programmers in industry: Programmers whose work may involve misinformation detection of texts may benefit from having an API-like tool like TRUExT, which can be an example of a tool that this stakeholder can use for accurate and quick misinformation detection for large databases of texts.
- Public tool users: TRUExT can also be used for open-source use by U.S. citizens who want to be informed on whether and why news articles that are found on mainstream news sources, independent news sources, or social media are trustworthy or not trustworthy.

[2]https://huggingface.co/

Therefore, the addition of TRUExT can contribute to a variety of applications including keeping military supply chains resilient, investment in the ongoing information war, and informing users on the veracity of a diversity of news articles.

What differentiates TRUExT from currently available data tool options is that most tools focus on fact-checking claims or evaluating social media posts and not as much on evaluating lengthy news articles. Likewise, if the tools do focus on evaluating news articles, there can be additional hurdles for its users, including paying subscription fees and/or requesting special access from the creators of these tools. TRUExT seeks to eliminate such barriers by making it open-source, akin to how open-source models are available for automatic download on the Hugging Face website.

II. Literature Review and Related Work

A. Infodemic Definitions and Stakeholder Applications

It is important to note which exact definition and form of misinformation needs to be considered in the context of the research problem for this study. First, there are specific distinctions between the following infodemic definitions [13]:

- Propaganda: Information shared by the government usually with a political connotation.
- Disinformation: False information that is intentionally disseminated to cause harm.
- Misinformation: False information that is not intentional in harm.
- Malinformation: False information that is deliberately distributed to cause harm to a person's or organization's reputation.

There are yet other types of fake news, including satire, hoax, clickbait, and conspiracy [22], as well as rumors and spam [16]. Opinionated news will be the target category for analysis in this study, as this is a form of intentional disinformation and is likely to have more easily recognizable NLP dimensions to embed including high sentiment, high exaggeration, and lack of context that can help predict the trustworthiness scores.

The intended stakeholders of U.S. government and military agencies, industry programmers, and public users in which this study could impact involve many applications. First, when it comes to the overall role of the U.S. military in providing defense for citizens, it is important that military agencies are able to meet supply and demand needs to make clear business decisions when building essential military equipment. Misinformation can impact supply chain risk management when false information about a company or country can jeopardize business decisions [1]. Next, when it comes to government agencies, this work could be applied to certain agencies (i.e., NSA, CIA) that focus on defending national security interests against misinformation threats, both foreign and domestic [13], [42]. In the context of programming in industry, not much research has been conducted on misinformation tools helping data scientists in industry, however easy-access, open-source tools such as APIs have been very helpful for industry programmers to quickly request information about

databases of entities on a large-data scale [44]. Finally, it is important for public users to have knowledge of and access to tools that can verify whether the news they see in various news media is trustworthy or not; current open-source fact-checking applications include Politifact, GossipCop, B.S. Detector, and Fake News Detector AI [33], [42]. Unlike these tools, TRUExT evaluates lengthy news articles from a wide diversity of sources instead of focusing on short statements from topic-centered sources.

B. Related Work in Communications Field

Numerous aspects of misinformation exist, both text-based and metadata. For example, the use of certain pronouns, politeness, emotion, complex vocabulary, and words indicating uncertainty can mean the difference between fake news and real news [4]. Characterizations of a higher likelihood of misinformation can include a distant speaker-audience relationship, high emotion, a goal to illicit audience action, detailed information, and a high number of participants and sources cited [5]. Other factors including diffusion scope (how broad the audience is), speed (number of users reacting within a certain timeframe), and shape (broadcast or human-human transmission) have been used to determine misinformation through analyzing clusters of users reacting to certain tweets on the social media network X, formerly known as Twitter [6]. Sentiment is regarded as an extremely important textual dimension of misinformation, and the use of emotion in misinformation can lead viewers to be more engaged and susceptible to misleading messages [21]. The textual analysis technique of NLP, which is where a computer is trained to interpret human language, is a method that can be used to incorporate communication dimensions into a data science model. This is usually done through converting words into numerical representations, also known as word embedding [40]. NLP tasks such as stance detection, rumor detection, and sentiment analysis can be performed by converting certain word vocabularies into numeric inputs [34]. Regarding integration of individual dimensions into misinformation detection tools, the TrustServista API has the capability to provide a holistic trustworthiness score and can take into account factors such as named entities, context (who/what/where), clickbait title, and sentiment, however it is acknowledged that there has not been much academic research conducted evaluating this API due to lack of transparency [35].

Considering which communication dimensions, to be translated into NLP dimensions, are most important for identifying what can be biased or opinionated news, a total of 13 dimensions stand out that can be identifiable using specified word embedding vocabularies. While not all of these dimensions are present in every news article, these dimensions should be universally considered when evaluating a news article for misinformation detection based on the literature analyzed in this study, and these dimensions will also be used as prediction factors for TRUExT. Dimensions were chosen in a way to not overlap much in indicative vocabularies but to cover broad accountability in determining trustworthiness.

1) Sentiment: Potentially the most characteristic dimension of subjective news [21], it emotionally impacts readers to believe in false messages in the article.
2) Persuasion: Similar to bias and opinion; language that attempts to get the reader to align with the article's point of view is a very likely indicator of misinformation [15].
3) Exaggeration: Exaggerated and/or outraged language is also indicative of misinformation [3].
4) Context: The more context that there is with cited persons of interest, the less likely that there is misinformation present in the article [2].
5) Inclusion of multiple perspectives: Multiple perspectives that take in different points of view from people of interest add more layers of context to the article and reduce bias [2].
6) Named entities: Named entities can be very important for adding more context to a news article and decreasing the likelihood of misinformation, especially if there is a greater diversity of named entities [10]. The less diversity in named entities present in an article, the more likely it is for misinformation to be present in the article.
7) Use of statistics: Statistics can be very effective tools for clarifying facts in a scientific manner and can add more context to a news article's message. Even mentions of dates and monetary figures can fall under the umbrella of statistics as these can clarify exact timeframes and figures. More statistics means more likelihood that a news article is trustworthy. This dimension is not commonly thought of when it comes to disinformation detection, however it is important enough to provide well-rounded context to an article [2].
8) Referencing of previous articles: As an alternative to cross-referencing, if the article references certain information from a previous news article or interview, or evidence from a previous report, that can add to the reliability of a news article [43].
9) Term frequency: If there are terms that are irregularly compared to words that are normally used, this can lead to more likelihood of misinformation [29].
10) Distraction: Words that indicate an attempt to distract or deflect from the main topic of the article can lead to a greater likelihood of misinformation [19].
11) Verification of claims: As an alternative to fact-checking, this dimension can provide factual clarity to claims that are made in a news article and add to the trustworthiness of the article [43].
12) Logical coherence: Words such as "first", "then", "therefore", and "consequentially" indicate a logical flow of ideas and subjects in a news article, which can add more context [43].
13) Clickbait title: Words present in the title that indicate clickbait or exaggeration can be indicative of an article not being trustworthy [25].

C. Transformer Model

The transformer model, in which LLM architectures would later be based on, was invented by Google in 2017 [38]. It is crucial to note that transformer models in their inherent training nature are not necessarily supervised. The original transformer model developed by Google was trained in a semi-supervised setting [38], meaning that it was trained on a small amount of labeled data but used the rest of the data that was unlabeled to make predictions. Similar methodologies were used for training the LLMs that will be discussed in this section. In the context of this study, the fine-tuned LLMs will be regarded as supervised learning models; although the models are pre-trained in a non-supervised setting, fine-tuning the LLMs to perform the intended NLP downstream tasks for this study will require labeled data and a supervised learning approach [14], [20].

D. Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers, also known as BERT, was first proposed by Google in 2018 [11]. This model was pre-trained using unlabeled data in an unsupervised setting to perform NLP tasks including masked language modeling (MLM) and next sentence prediction [11]. BERT utilizes the original transformer architecture [38], where the input embeddings for the model are the sum of the token embeddings, the segmentation embeddings, and the position embeddings [11] that are encoded from the text using the BERT tokenizer. It also incorporated significant improvements upon previous language representation models, which had flaws of unidirectional NLP tasks and left-to-right context evaluation [11]. BERT's ability to fuse the left and right (bidirectional) contexts of sentences together to evaluate text holistically, as well as its versatility in doing text prediction, question answering, and text classification [11], made it a powerful tool for various NLP tasks.

E. Robustly Optimized BERT Approach

Robustly Optimized BERT Approach, also known as RoBERTa, was proposed by Facebook AI Research in 2019 as an improvement upon the original BERT model [20]. The model was pre-trained using unlabeled data but fine-tuned using labeled data [20], which is also known as self-supervised learning. As BERT had certain performance training flaws in need of improvement, the upgrades that RoBERTa incorporated included training the model longer, using larger batches and sequences of textual data, and dynamically changing the masking pattern applied to the training data [20]. Incorporating these changes showed that the accuracy and performance of NLP downstream tasks on various open-source datasets improved significantly using RoBERTa compared to BERT, as the longer the model was trained, the more these results reflected that performance [20].

F. Distilled BERT

Since BERT was trained on a larger number of parameters and required a large amount of data to be trained on, it was computationally expensive to train and perform NLP tasks, therefore DistilBERT was introduced as a computationally faster and lightweight alternative [30]. DistilBERT is 40% smaller than BERT and retains 97% of BERT's performance; its accuracy in performing downstream NLP tasks is similar to BERT's [30]. This model architecture functions by containing the same multi-head attention and feed forward mechanisms yet incorporating fewer transformer layers than BERT-base's original layers while preserving BERT-base's efficacy [30].

G. Related Work in Data Science Field

Many supervised machine learning and deep learning models have been used for misinformation detection tasks including Naive-Bayes classifier, support vector machine, random forest, XGBoost, recurrent neural networks (RNNs), deep belief networks, and convolutional restricted Boltzmann machines [16], [28]. Additionally, text pre-processing mechanisms including regular expressions, lemmatization, stop word removal, conversion to N-gram vectors and term frequency-inverse document frequency (TF-IDF) sequence vectors have been used for converting text into vectorized inputs for deep learning models imported for fake news detection [18]. A Long Short Term Memory Recurrent Neural Network (LSTM-RNN) was used for classifying misinformation on a binary scale (e.g., an output of 0 for misinformation or 1 for not misinformation) [8]; classification tasks have been preferred compared to regression tasks in misinformation detection. Different types of misinformation data for detection tasks have also been utilized in field research [24], [31], [36].

Misinformation is likewise a global problem; while detecting it in most common languages like English is well studied, it is often difficult to identify in less commonly used languages due to the presence of the foreign-language effect obscuring intuition when it comes to determining misinformation potential [37].

BERT, RoBERTa, DistilBERT, and many other LLMs are available for open-source use on the Hugging Face website [41]. These models can be downloaded using the Transformers package in Python and fine-tuned to perform any NLP task including chat generation, sentence prediction, text classification, and text summarization. Fine-tuned models designed to perform specific NLP tasks can also be deployed to the website. Generally, the maximum number of input tokens that these models can process is 512 [11], which is equivalent to around 400 words. LLMs have been used for misinformation classification tasks on textual data, and these transformer models have overall provided good results for performing these tasks [7], [17], [27].

H. Research Gaps

Although much research has been conducted using advanced machine learning models and LLMs for misinformation detection, there are still some significant research gaps to note. First, more tools are still needed for automated evaluation of trustworthiness in news sources, in addition to a lack of explainability in current tools [26]. Thus, it is important

to contribute another tool to the range of available options while also adding significance to the area of XAI, a topic that is currently garnering much attention in the field of data science. Additionally, supervised learning classification tasks in misinformation detection are more common than regression tasks, so there needs to be an exploration of predicting cumulative trustworthiness scores (preferably between 0 and 100) through the development of TRUExT. Lastly, Hugging Face LLMs take a maximum of 512 tokens, which unfortunately means for lengthy article texts that are longer than 400 words, these LLMs cannot capture the entire context of these texts. Therefore, there need to be additional measures implemented as inputs into the LLMs to improve score predictions while holistically evaluating the article texts, which the integration of the communication-based NLP dimensions will attempt to compensate for this gap.

III. Methodology

A. Data Collection

A total of 50,092 English-language news articles, from approximately 1,555 unique article origins, with a median length of 485 words per article, were compiled between several news datasets to create the training dataset [12] (see list below). Some of the larger datasets had random samples of articles taken for the final compilation so there would not be an excessive number of articles in the starting dataset between the multiple data sources; approximately 50,000 articles was a large enough dataset for the initial training of the LLM. To make compilation easier these datasets had to have fields of article origin (i.e., publication, URL, domain), article title, and article text. The article title and article text were the most important fields as these contained the necessary NLP dimensions to embed for analysis and prediction. The article origin was used for assessment of news source diversity, however when creating TRUExT the ultimate objective is for the tool to detect misinformation requiring as few fields as possible and to make predictions based only on the text and not on external factors like article origin, article author, etc. If the tool required additional fields, it should consequently be considered that potential users of TRUExT may have difficulty retrieving that information, thereby leading to incomplete or inaccurate predictions from the base model.

The news article data originate from the following Kaggle and GitHub sources with the content scope for each source listed below:

- In-house project dataset (webscraped articles and blog posts) [12]
- All the News Kaggle dataset (mainstream and independent news)
- MC-Fake dataset (news on politics, entertainment, health, COVID-19, and war in Syria) [23]
- FakeNewsCorpus dataset (independent news categorized as satire, extreme bias, conspiracy theory, hate news, etc.)
- Fake News Detection Kaggle and Getting Real about Fake News Kaggle datasets (combination of real and fake news, fake news articles extracted using webhose.io API)
- FakeCovid dataset (COVID-19 news) [32]

B. Data Preprocessing

For data preprocessing, most textual context was retained as the LLMs are designed to capture the bidirectional contexts of sentences, therefore the news articles did not require text cleaning steps such as stopword removal, lemmatization, etc. as this would lose context for the LLM trainings.

Indicative vocabularies for the 13 communication-based dimensions (see Section II.B.) were defined. The number of times each dimension vocabulary token appeared in each tokenized article text would then be counted. Next, each dimension count would be divided by the length of the article text to normalize the proportion of indicative vocabulary appearances in the text. See [12] for more details and figures on the spread and shape of each normalized dimension count distribution. Using percentile thresholds of the distributions for these normalized dimension counts, each dimension would be assigned a weight between 0 and 10 depending on the presence or non-presence of each dimension that would contribute to a greater or lesser likelihood of misinformation. The dimensions referencing of previous articles (2), distraction (2), clickbait title (2), and verification of claims (4) were exceptions to having a maximum weight of 10 assigned due to not having much variability in their normalized count distributions and inconsistent presence in the news articles [12]. To justify, if these 4 dimensions each had a maximum weight of 10, because they were not consistently present in every news article the labeled cumulative scores would be much lower due to a factor that would not be present in most articles. Therefore, adding up the dimension weights (9 with a maximum of 10, 3 with a maximum of 2, and 1 with a maximum of 4), would result in a cumulative score between 0 and 100 to be suitable for the LLM regression training task, which is how the news article texts were labeled for training.

C. LLM Training

To set up each LLM training, the quantified NLP dimensions for each given article text were concatenated with the text to improve the prediction accuracy of the LLM. The training dataset was split up into 70% training, 15% validation, and 15% testing. All LLMs were imported from the Transformers package in Python and were trained using the TensorFlow package.

For all LLM trainings, a Keras Dense layer was utilized as a regression head to fine-tune the model and generate target outputs. The hyperparameter combination used for all trainings, including an Adam optimizer, a learning rate of $1e^{-5}$, and a batch size of 32, was consistently chosen to have fast and efficient training of the LLMs while keeping training stable to avoid rapid loss function convergence. Mean squared error (MSE), mean absolute error (MAE), Pearson Correlation Coefficient, Spearman Rank Correlation Coefficient, and R^2 were used as evaluation metrics. The imported Hugging Face models from the Transformers package that were trained

and compared were "bert-base-uncased", "roberta-base", and "distilbert-base-uncased".

D. Results

Without integration of the NLP dimensions, the accuracy for the LLM predictions would be approximately 0.53 and the correlation coefficient would be around 0.73 [12]. Conversely, all 3 LLMs that were trained improved greatly in accuracy and goodness of fit with the addition of the NLP dimensions as input, and the RoBERTa model training resulted in the lowest MSE and highest overall correlation in trustworthiness score predictions (see Tables I, II, and III). However some flaws to note were that RoBERTa took in the lowest amount of maximum input tokens (at 160, as was allowable by the Google Colab Pro GPU), and required the longest training time. The BERT model training resulted in the highest MSE and lowest R^2 while the DistilBERT training allowed for the highest amount of maximum input tokens (see Tables I, II, and III).

TABLE I
MODEL TRAINING RESULTS

Model	***Number of Epochs***	***Maximum Input Tokens***	***Training Loss (MSE)***	***Training MAE***
BERT	8	176	4.4780	1.4970
RoBERTa	17	160	**1.6063**	**0.9956**
DistilBERT	12	300	3.6656	1.4519

TABLE II
MODEL VALIDATION AND TESTING RESULTS

Model	***Validation Loss (MSE)***	***Validation MAE***	***Testing Loss (MSE)***	***Testing MAE***
BERT	2.8110	1.1227	2.7812	1.1202
RoBERTa	**2.2895**	1.1895	**2.2804**	1.2012
DistilBERT	2.3013	**1.1119**	2.2903	**1.1128**

TABLE III
MODEL CORRELATION COEFFICIENT RESULTS

Model	***Pearson Coeff.***	***Spearman Rank Coeff.***	R^2 ***(Pearson)***	R^2 ***(Spearman Rank)***
BERT	0.9884	0.9942	0.9770	0.9885
RoBERTa	**0.9952**	**0.9966**	**0.9904**	**0.9932**
DistilBERT	0.9903	0.9928	0.9807	0.9856

In addition, statistical significance tests were performed between combinations of the model scores and the labeled scores to validate if the RoBERTa model would still be the best choice for the base model for TRUExT using a subset of 1000 observations from the training dataset. The results shown in Table IV indicate that all sets of prediction scores using the LLMs and the labeled scores were significantly different. The RoBERTa prediction scores also had the closest distribution mean compared to the actual labeled scores' distribution mean, thereby validating it as the superior choice for the model foundation for TRUExT.

TABLE IV
STATISTICAL SIGNIFICANCE TEST RESULTS

T-test	***T-statistic***	***P-value***
BERT & RoBERTa	-37.7539	$6.5889e^{-236}$
RoBERTa & DistilBERT	-6.0106	$2.1912e^{-9}$
BERT & DistilBERT	-40.5685	$5.3107e^{-263}$
Real Scores & BERT	17.7735	$9.9483e^{-66}$
Real Scores & RoBERTa	-3.2059	0.0014
Real Scores & DistilBERT	-6.7062	$2.5910e^{-11}$
ANOVA	**F-statistic**	**P-value**
BERT, RoBERTa, & DistilBERT	1000.9154	0

Further validation of these results was done by using a subset of the WELFake Kaggle dataset [39]. Using this subset, the scores that would be calculated using the NLP dimension labeling algorithm versus the scores that would be calculated using the RoBERTa model were compared with the labels in the WELFake dataset. Note that the labels in the WELFake dataset were binary (0 and 1), so only classification accuracy could be used to compare these results. As shown in Table V, although the accuracies were low due to the NLP dimension vocabularies lacking complexity and real-world examples found in the news today (implementing these examples would likely improve the model's efficacy in the future and this limitation will be elaborated upon in Section IV), these results show that the RoBERTa model still performed better than the NLP dimension labeled score calculations. These predictions were also statistically significant as the T-test between these sets of predictions resulted in a T-statistic of -20.345 and a p-value of $1.180e^{-83}$.

TABLE V
NLP DIMENSION PREDICTIONS VS. LLM PREDICTIONS

Type of Predictions	**Accuracy**	**Precision**	**Recall**	**F1-Score**
NLP Dimensions	0.3011	0.2475	0.2070	0.2255
RoBERTa	**0.4486**	**0.4612**	**0.7267**	**0.5643**

E. TRUExT Architecture and RoBERTa Integration

TRUExT, as given by its proposed name, was designed to be a regression-based trustworthiness score predictor for news articles originating from a wide diversity of sources while unifying, or integrating, the 13 specified communication dimensions of misinformation as NLP features with the chosen RoBERTa model based on the LLM comparison results. Combining the functions of these two crucial components into TRUExT, the NLP dimensions likewise serve as explainability measures to justify why the LLM trustworthiness regressor provided the score that it would give an article text based on the presence or non-presence of each dimension, therefore creating a level of transparency for potential users of TRUExT.

TRUExT would require an input dataframe of article texts and titles to generate effective predictions. To integrate the

RoBERTa model into TRUExT, the technique of extracting the quantified NLP dimensions and concatenating them with the article texts would be performed first to create expected inputs for the tool. The concatenated texts would be encoded using the RoBERTa tokenizer and converted into input IDs and attention masks to feed into the RoBERTa base model. Note that the exact model architecture and weights would be replicated as it was during training in order for the model prediction to function properly. As the LLM would then generate each score prediction, in addition to the prediction TRUExT would output explainability metrics (shown in Fig. 2) to explain why the model likely gave the predicted score to the given article text. Finally, the predicted scores would be stored in a new column in the input dataframe and the dataframe with the predictions would be exported to the user. Fig. 1 shows an outline of the structure of TRUExT with its proposed processing of inputs and outputs.

IV. Discussion

A. Scientific and Societal Impacts

The findings from this study proposing TRUExT have shown to be significant and can pose major contributions to the field of data science as well as to interdisciplinary research in the misinformation domain. As there are currently very few commercially available tools that utilize communication-based dimensions to predict a regression-based score for news articles, TRUExT should serve as a starting contribution for data tools specifically designed with these objectives. Out of previously known options, the TrustServista API did accomplish this but as previously mentioned, not much research has been conducted using this API due to a lack of transparency with its host company Zetta Cloud [35]. TRUExT was able to integrate 13 communication-based NLP dimensions of misinformation with a Hugging Face LLM regressor to produce explainable trustworthiness score predictions for a variety of news articles, making it a tool that can build on the lack of available misinformation detection tools that are usually used for social media data, fact-checking, etc. while also filling in crucial gaps on XAI and transparency in misinformation detection models. Out of the 3 LLMs tested during the course of this study, RoBERTa performed the best therefore it would still function as a very effective and practical base model for TRUExT due to its strong accuracy results.

Based on the variety of literature review for this study, 13 communication dimensions of misinformation were proposed for integration in TRUExT. While some dimensions may not always be present in news articles, all should be regarded as individual holistic measures of trustworthiness to determine whether a news article is trustworthy or not. This view can also contribute a unique perspective to the field of communications and journalism on which textual dimensions of misinformation are most crucial to examine and consider when making predictions, thereby circumventing metadata factors such as author, publication credibility, etc. that may be more difficult to obtain and utilize in model predictions. In general, communication dimensions converted into NLP features can be a recommended approach to improving the accuracy of future misinformation detection models and transdisciplinary research.

The NLP dimensions served an extremely significant and versatile role in labeling data, quantifying communication dimensions of misinformation, improving model accuracy, and providing explainability in the tool, therefore the addition of these features helped immensely in providing quality to the base model predictions. Likewise, the addition of the NLP dimensions contributed an effective method for bypassing the limitation of the LLMs requiring a limited number of input tokens while still making strongly accurate predictions, thereby filling another critical research gap outlined in Section II.H.. The chosen base LLM of RoBERTa for TRUExT required 160 input tokens (even less than the maximum 512) as was allowable using Google Colab Pro GPU resources. However, TRUExT's ability to simultaneously incorporate the extracted quantified NLP dimensions into the score predictions created an effective data science approach to comprehensively evaluate the entire content of news articles despite the LLM input token limitation.

In Section I, some target societal stakeholders were introduced that this study would impact, which included U.S. government agencies, U.S. military agencies, industry programmers, and public users. These target audiences can greatly benefit as there is much potential for them to utilize TRUExT for rapid and automated prediction of comprehensive trustworthiness scores for a wide diversity of news articles, with added transparency to improve the trust of this technology with potential users.

The contribution of TRUExT could likewise be potentially instrumental in information warfare by government agencies that may currently be investing in tactics against foreign and domestic adversaries that spread harmful misinformation; this tool can provide an opportunity for them to advance in the ongoing information war. Military agencies can also use TRUExT to combat adversaries that may spread defaming misinformation about potential suppliers crucial for supply chain management for the U.S. military. Thus, having a truthful image about which suppliers are trustworthy to meet supply needs for the military is important to keep supply chain operations and military readiness consistently strong. For industry programmers, the tool can also help these programmers rapidly and accurately populate trustworthiness scores for databases of news articles, helping facilitate misinformation detection tasks in industry. Public users of TRUExT can be informed effectively on which news, originating from most types of media, are likely to be misinformation or not, potentially serving as a contribution to improving media literacy. Overall, the societal impacts of this study included that TRUExT provided a transparent, big-data solution to help protect society against the harmful effects of misinformation through more accurate trustworthiness scoring of news articles while being an easy-access, open-source option for a crucial target audience of potential users.

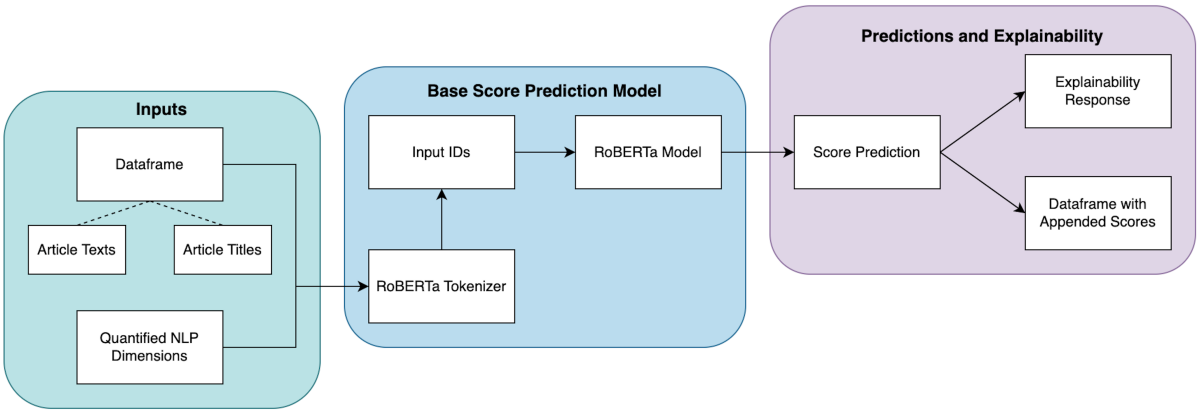

Fig. 1. Tool Structure of TRUExT


```
{
    "trustworthiness_score": 67,
    "dimension_scores": {
        "sentiment": 9,
        "persuasion": 7,
        "exaggeration": 8,
        "context": 4,
        "multiple_perspectives": 6,
        "named_entities": 10,
        "statistics": 2,
        "previous_articles": 1,
        "term_frequency": 3,
        "distraction": 2,
        "claim_verifications": 3,
        "logical_coherence": 10,
        "clickbait_title": 2
    },
    "explainability": {
        "sentiment": "not present",
        "persuasion": "little presence",
        "exaggeration": "little presence",
        "context": "some presence",
        "multiple_perspectives": "some presence",
        "named_entities": "present",
        "statistics": "not present",
        "previous_articles": "some presence",
        "term_frequency": "little presence",
        "distraction": "not present",
        "claim_verifications": "present",
        "logical_coherence": "present",
        "clickbait_title": "not present"
    }
}
```


Fig. 2. Sample TRUExT Response

B. Limitations

Despite many potential contributions to the misinformation detection domain, there are some limitations of this proposed tool that need to be noted. The most significant limitation of this study that can impact TRUExT's peer review validation for deployment in the future is that the vocabularies used for encoding the NLP dimensions were vague in some qualities. None of the dimension vocabularies had indicative phrases spanning more than one word. Since tokenization only involves single words or punctuation marks, while it is easy to count the number of tokens that match the vocabulary words on a single-word basis, tokenization currently limits vocabulary match encoding for phrases, and in future research this must not impose a major limitation on being able to quantify the NLP dimensions authentically. Dimension vocabularies also need to be specified, real-world examples found in current news articles.

Since the distribution of the normalized dimension count trustworthiness scores followed a normal distribution, there were many articles that had a trustworthiness score between 45 and 55, meaning that there were many articles that were classified as likely trustworthy (meaning a score of 50 or above) or untrustworthy (meaning a score below 50) based on a few-point difference, therefore the NLP dimension weights may have to be adjusted more in the future to fit realistic expectations [12].

There are also limitations regarding the overall functionality of TRUExT. First, the framework for quantifying the NLP dimensions is mandatory for the LLM and the tool to function properly and predict accurately. Without the inclusion of this NLP framework, the LLM would severely underperform in accuracy [12], thus the trustworthiness score predictions would be highly inaccurate if the predictions were solely dependent on the short truncated texts passed through the model. Next, TRUExT only requires the fields of article text and article title in order to make predictions. The ultimate objective of this study was to devise a tool that could evaluate misinformation potential solely on text and to require as few fields as possible. More fields including article author, article publication, and other outside factors could potentially be difficult for users of TRUExT to collect, however it should be noted that these factors can make a supplemental impact on whether a news article is likely to be trustworthy or not.

TRUExT does not work using non-English-language news articles, as the dimension vocabularies contained only English words and punctuation. It is important to note that misinformation detection in non-English languages is still a major research gap in the misinformation domain [37], however this work will not be able to contribute to this gap as compiling

NLP dimension vocabularies in non-English languages would require extensive time and effort and likely a translation API that would add an additional monetary cost to research.

Next, the base LLM for TRUExT was trained on only around 50,000 news articles. Although these news articles were from a wide diversity of sources, it is possible that the training dataset may have to be expanded to include 100,000 or even 200,000 news articles to ensure consistent predictions. As the news articles in the training dataset become less recent, more recent news articles will have to be added over time to include current information or instances of misinformation that have been verified; the NLP dimension vocabularies will also require these same updates in the future.

A final point of discussion regarding this study and the proposal of TRUExT is that different individuals can have differing opinions on what news is considered misinformation or not. Humans in the loop are still crucial to analyzing misinformation from an objective lens, however some people may tend to classify news articles as likely not to be misinformation that may be, for example, noticeably higher in sentiment or bias. Potentially, some people who believe in conspiracy theories as truth may label a training dataset of news articles much differently compared to people who would try to evaluate the trustworthiness of news articles with as little personal opinion as possible. Therefore, different interpretations of what constitutes misinformation, based on individuals' subjective views of information, can impact whether these individuals would be able to find an automated tool like TRUExT trustworthy to use due to personal disagreements on what the tool might classify as misinformation or not misinformation.

V. Future Work

For future work, expanding and improving the complexities of the NLP dimension vocabularies is a limitation that can be greatly enhanced. First, phrases will need to be added to the vocabularies to specify examples of misinformation that could otherwise be misinterpreted if only spanning a single word (i.e., "steal the election" versus "steal"). NLTK tokenization should be eliminated and instead string mentions should be used for counting the instances of dimension vocabulary matches. Lemmatization could also be used to match lemmatized versions of vocabulary phrases to reduce the effort of vocabulary compilation. Next, inclusion of vocabularies that are not indicative of the dimensions could also be added to see if the news articles contain, for example, more words indicating negative sentiment versus neutral sentiment, or more words indicating objective views versus biased views. However, a major limitation to improving the vocabularies to recognize is that expanding these to cover a wide variety of indicative words and phrases can take a lot of time and will have to consistently be updated with current events and terminology as more common instances of misinformation appear over time, yet it is crucial to emphasize that expanding the dimension vocabularies would greatly improve upon the quality and reliability of this work.

Next, since the distribution of the labeled scores followed a normal shape, making the distribution follow a more bimodal shape through adjusting the dimension weights could help classification tasks be more accurate in the future given a cumulative score threshold, therefore there is more work that can be done to re-evaluate the proper weights of the NLP dimensions. While not all news articles contain all 13 dimensions examined in this study, and some were weighted less due to these not being consistently present in news articles, cumulative scores could still be calculated depending on whether a news article has met at least a certain number of dimension requirements.

The base LLM in TRUExT can be trained on more data, even if this will require more training time. Additional news articles from other open-source datasets would be helpful to include even more variety in article origins, and potentially non-English-language news articles could be incorporated to help close the gap on non-English-language misinformation detection. However, for TRUExT to help fill this gap, NLP dimension vocabularies in non-English languages will have to be added, which this expansion could require extensive compilation and a translation API that would mean additional monetary costs.

VI. Conclusion

While the ideal solution to the problem of misinformation is widespread media literacy and legislative regulations, current big-data solutions require more options that cover gaps on explainability and trustworthiness predictions of news articles. Target stakeholders likewise need to have awareness that such tools exist in order to be educated on which news is true or fake, with technological accountability to establish trust in these tools. Despite the harrowing task of winning the information war, data science has a major role to play and TRUExT, through its introduction in this study, could have the potential to become a widely-used tool by government agencies and open-source users alike for accurately and effectively informing on whether certain articles found in the news are likely to be misinformation or not, given that the limitations of this work will be improved upon later.

Acknowledgment

The authors would like to acknowledge Dean Hougen, Naveen Kumar, Katerina Tsetsura, and Jeong-Nam Kim for their steadfast support and guidance with this interdisciplinary study.

References

[1] Akhtar, P., Ghouri, A. M., Khan, H. U. R., Amin ul Haq, M., Awan, U., Zahoor, N., ... & Ashraf, A. (2023). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. *Annals of Operations Research, 327*(2), 633-657.

[2] Amer, E., Kwak, K. S., & El-Sappagh, S. (2022). Context-based fake news detection model relying on deep learning models. *Electronics, 11*(8), 1255.

[3] Bhat, M. M. (2022). *Study of Effectiveness of Stylometry in Misinformation Detection* (Doctoral dissertation, The Ohio State University).

[4] Chiu, M. M., Morakhovski, A., Ebert, D., Reinert, A., & Snyder, L. S. (2023). Detecting COVID-19 fake news on Twitter: Followers, emotions, relationships, and uncertainty. *American Behavioral Scientist*, 00027642231174329.

[5] Chiu, M. M., & Oh, Y. W. (2021). How fake news differs from personal lies. *American behavioral scientist, 65*(2), 243-258.

[6] Chiu, M. M., Park, C. H., Lee, H., Oh, Y. W., & Kim, J. N. (2022). Election Fraud and Misinformation on Twitter: Author, Cluster, and Message Antecedents. *Media and Communication, 10*(2), 66-80.

[7] Cui, Z. (2022, January). COVID-19 Fake News and Misinformation Detection using Transformer Learning. In *2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM)* (pp. 965-968).

[8] Cunha, B., & Manikonda, L. (2022). Classification of Misinformation in New Articles using Natural Language Processing and a Recurrent Neural Network. *arXiv preprint arXiv:2210.13534*.

[9] Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent arts & humanities, 9*(1), 2037229.

[10] De Magistris, G., Russo, S., Roma, P., Starczewski, J. T., & Napoli, C. (2022). An explainable fake news detector based on named entity recognition and stance classification applied to covid-19. *Information, 13* (3), 137.

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

[12] Donner, C. G. G. (2024). *Misinformation Detection Methods Using Large Language Models and Evaluation of Application Programming Interfaces*. [Master's thesis, University of Oklahoma]. SHAREOK Repository. https://hdl.handle.net/11244/340251

[13] Gradoń, K. T., Hołyst, J. A., Moy, W. R., Sienkiewicz, J., & Suchecki, K. (2021). Countering misinformation: A multidisciplinary approach. *Big Data & Society, 8*(1), 20539517211013848.

[14] Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

[15] Hameleers, M., & Brosius, A. (2022). You are wrong because I am right! The perceived causes and ideological biases of misinformation beliefs. *International Journal of Public Opinion Research, 34*(1), edab028.

[16] Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining, 10*, 1-20.

[17] Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences, 9*(19), 4062.

[18] Kong, S. H., Tan, L. M., Gan, K. H., & Samsudin, N. H. (2020, April). Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)* (pp. 102-107). IEEE.

[19] Kwek, A., Peh, L., Tan, J., & Lee, J. X. (2023). Distractions, analytical thinking and falling for fake news: A survey of psychological factors. *Humanities and Social Sciences Communications, 10*(1), 1-12.

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[21] Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications, 5*, 1-20.

[22] Mhatre, S., & Masurkar, A. (2021, June). A hybrid method for fake news detection using cosine similarity scores. In *2021 International Conference on Communication information and Computing Technology (ICCICT)* (pp. 1-6). IEEE.

[23] Min, E., Rong, Y., Bian, Y., Xu, T., Zhao, P., Huang, J., & Ananiadou, S. (2022). Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022* (pp. 1148-1158).

[24] Oh, Y. W., & Park, C. H. (2021). Machine cleaning of online opinion spam: Developing a machine-learning algorithm for detecting deceptive comments. *American behavioral scientist, 65*(2), 389-403.

[25] Oliva, C., Palacio-Marín, I., Lago-Fernández, L. F., & Arroyo, D. (2022, August). Rumor and clickbait detection by combining information divergence measures and deep learning techniques. In *Proceedings of the 17th International Conference on Availability, Reliability and Security* (pp. 1-6).

[26] Przybyła, P., & Soto, A. J. (2021). When classification accuracy is not enough: Explaining news credibility assessment. *Information Processing & Management, 58*(5), 102653.

[27] Rai, N., Kumar, D., Kaushik, N., Raj, C., & Ali, A. (2022). Fake News Classification using transformer based enhanced LSTM and BERT. *International Journal of Cognitive Computing in Engineering, 3*, 98-105.

[28] Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems, 34*(2), 76-81.

[29] Riego, N. C. R., & Villarba, D. B. (2023). Utilization of Multinomial Naive Bayes Algorithm and Term Frequency Inverse Document Frequency (TF-IDF Vectorizer) in Checking the Credibility of News Tweet in the Philippines. *arXiv preprint arXiv:2306.00018*.

[30] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[31] Serrano, J. C. M., Papakyriakopoulos, O., & Hegelich, S. (2020, July). NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

[32] Shahi, G. K., & Nandini, D. (2020). FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*.

[33] Sharma, D. K., Garg, S., & Shrivastava, P. (2021, February). Evaluation of tools and extension for fake news detection. In *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)* (pp. 227-232).

[34] Su, Q., Wan, M., Liu, X., & Huang, C. R. (2020). Motivations, methods and metrics of misinformation detection: an NLP perspective. *Natural Language Processing Research, 1*(1-2), 1-13.

[35] Titiliuc, C., Ruseti, S., & Dascalu, M. (2020, September). What's Been Happening in the Romanian News Landscape? A Detailed Analysis Grounded in Natural Language Processing Techniques. In *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (pp. 195-201). IEEE.

[36] Turner, J., Kantardzic, M., Vickers-Smith, R., & Brown, A. G. (2023). Detecting Tweets Containing Cannabidiol-Related COVID-19 Misinformation Using Transformer Language Models and Warning Letters From Food and Drug Administration: Content Analysis and Identification. *JMIR infodemiology, 3* (1), e38390.

[37] van de Meerakker, K. O. (2022). The foreign language effect on the credibility of fake news messages among Dutch news readers and the influence of emotional loading.

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[39] Verma, P. K., Agrawal, P., Amorim, I., & Prodan, R. (2021). WELFake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems, 8*(4), 881-893.

[40] Wadud, M. A. H., Mridha, M. F., & Rahman, M. M. (2022). Word embedding methods for word representation in deep learning for natural language processing. *Iraqi Journal of Science*, 1349-1361.

[41] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[42] Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. *European Journal of Operational Research, 279*(3), 1036-1052.

[43] Zhang, X. Examining COVID-19 Vaccination Misinformation and Clarification by the Public Sector in Hong Kong.

[44] Zhong, H., & Mei, H. (2017). An empirical study on API usages. *IEEE Transactions on Software Engineering, 45*(4), 319-334.