Categorising Corruption in the Vaccine Discourse: A General Taxonomy, Data Set, and Evaluation of LLMs for Classifying Corruption Dialogue in Social Media

Vitor Gaboardi dos Santos[1], Guto Leoni Santos[1], Antonia Egli[1], Estatira Kahvazadeh[2], Bill Doolin[3], Patricia Takako Endo[4], and Theo Lynn[1]

[1] Dublin City University, Ireland
{vitorgaboardidos.santos,guto.santos,antonia.egli,theo.lynn}@dcu.ie
[2] Georgia Institute of Technology, United States ekahvaz@gmail.com
[3] Trinity College Dublin, Ireland doolinbill@gmail.com
[4] Universidade de Pernambuco, Brazil patricia.endo@upe.br

Abstract. Real or perceived corruption can have a damaging effect on health care services and outcomes. In particular, research suggests perceived corruption had a significant impact on COVID-19 vaccination. Given the role of social media in health communications, identifying and understanding perceived corruption related to vaccines and vaccination is critical to build societal cohesion and public trust in health institutions and strategies, manage and combat misinformation and disinformation, and design more effective policies, interventions, and communications strategies. There is a dearth of research on binary and multi-class classification of corruption dialogues in health or otherwise. We address this gap by introducing a general hierarchical corruption dialogue taxonomy (HCDT) and formulating binary and multi-class classification tasks based on the HCDT. We also create a vaccine-specific labelled dataset for each task, and fine-tune three large language models (BERT, RoBERTa, and BERTweet) based on these datasets. We evaluate the performance of these models in the binary and multi-class classification tasks. While all models performed similarly for the binary task, RoBERTa performed best for multi-class classification of corruption dialogue.

Keywords: Corruption · Large Language Models · BERT · Twitter · Multi-class Classification · Vaccine · COVID-19

1 Introduction

As of April 2024, over 775 million cases and 7 million deaths resulting from COVID-19 were reported to the World Health Organisation [1]. While a wide range of countermeasures were implemented to mitigate the spread of the disease, immunisation is the primary response against severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2), particularly as it continues to evolve [2]. Unfortunately, there still remains a significant population who are either vaccine hesitant or opposed to vaccination for COVID-19 or in general. A variety of factors contribute to such beliefs and behaviours including perceptions of trust in the vaccine approval process and vaccine effectiveness in protecting individuals, vaccine conspiracy beliefs, perceived side effects, perceived availability, and free-riding based on herd immunity [3]. The consequences are significant. Research suggests that between 2021 and 2022, over 232,000 deaths could have been prevented among unvaccinated adults in the United States alone [4, 5].

Corruption is commonly defined as “the abuse of entrusted power for private gain" [6, 7]. Corruption, real or perceived, can have a damaging impact on health outcomes and the quality of health care services and result in higher healthcare costs, erosion of trust in the health system, and reduced service utilisation [8]. In the context of COVID-19, a study of 90 countries worldwide found that public corruption was one significant causes of cross-country variation in immunisation progress [9]. The speed, scale, and complexity of approving, allocating, distributing, and rolling out COVID-19 vaccines worldwide was unprecedented and as a result introduced new corruption risks [10–12]. These risks include corruption in the vaccine development and approval process, vaccine deployment and distribution systems, vaccine procurement, emergency funding for vaccines, preferential access to vaccines, and vaccine policy decisions [10]. While social media has many benefits, it can interfere with public health communication by spreading health misinformation and disinformation, creating a false sense of uniformity and validity, and legitimising questionable or false information [13, 14]. This is particularly true in the context of the COVID-19 pandemic, where numerous studies highlight the adverse impact of social media on vaccine uptake [15–17]. Consequently, identifying and understanding perceived corruption related to vaccines and vaccination is critical to build social cohesion and public trust in health institutions and strategies, manage and combat misinformation and disinformation, and design more effective policies, interventions, and communications strategies.

In this paper, we propose a pipeline for automatically detecting and classifying perceived corruption in the vaccine discourse on social media platforms using Large Language Models (LLMs). While there is an extensive literature on corruption in health contexts, and specifically COVID-19, there are few studies on the use of Machine Learning (ML) to classify content for corruption-related dialogue on social media and vaccination. This can be explained by a number of significant challenges, not least a lack of corruption dictionaries and taxonomies, access to data both for testing and training models and empirical analysis, and availability of classification tools to identify posts relating to corruption.

In summary, we make four contributions in this paper to advance research on corruption dialogue detection and classification. First, we create a hierarchical corruption dialogue taxonomy (HCDT) that provides a structured framework for categorizing various forms of corrupt practices. Second, we create a labelled corruption dialogue dataset comprising tweets related to corruption within the COVID-19 vaccine discourse using a combined approach with humans and GPT [18]. Third, we fine-tune different LLMs using BERT-based architectures to initially detect corruption-related tweets and then categorize between

11 forms of corruption practices. Fourth, we evaluate the fine-tuned LLMs and report on their effectiveness in detecting and categorising binary and multi-class corruption-related tweets accurately. To the best of our knowledge, this is the first time an annotation scheme and computational model for identifying and classifying corruption-related discourse on social media has been designed and evaluated.

The remainder of this paper is organised as follows: Section 2 introduces related works on the detection of corruption using ML techniques. Section 3 describes our hierarchical taxonomy for corruption dialogue. Section 4 presents the data and methodology used to fine-tune and evaluate the LLM employed to detect and classify corruption dialogue. Section 5 discuss the results of our methodology. Section 6 describes limitations and avenues for future work. Finally, Section 7 presents the conclusion and final remarks of the paper.

2 Related Work

Several studies have been published on the use of ML techniques to detect corruption or related indicators in different domains. For example, Lima & Delen [19] conducted a study to predict levels of corruption perception indexes across 132 nations. They employed different ML models and used data collected from various website sources associated with indexes. The results revealed that the Random Forest (RF) model performed better, achieving an accuracy of 85.77%.

Rabuzin & Modrušan [20] compared different models to identify suspicious one bid tenders, which may raise suspicions of favouritism, collusion, or lack of transparency. They found that the Logistic Regression (LR) model produced the best accuracy overall, while Naïve Bayes (NB) exhibited the best performance in identifying potential signs of corruption in the public procurement process. While they found a lack of data within the field, they noted that the models show promise for public procurement corruption detection.

Denisova-Schmidt et al. [21] examined anti-corruption education effects on students' perceptions of academic integrity and corruption in Russia by employing a two-step ML regression process analysis using around 2,000 surveys. They found out that students who plagiarise frequently seem to have more negative opinions regarding corruption, suggesting that policymakers should consider unwanted diverse impacts across student groups before implementing anti-corruption education on a greater scale.

Ash et al. [22] used ML techniques to identify instances of corruption in local governments by analysing budgetary data from Brazilian municipalities. They employed annual budget data spanning from 2001 to 2012 to train models aimed at predicting the occurrence of corruption within these municipalities, achieving an accuracy rate of 76%. Their findings suggest that even in areas where corruption is not prevalent, there may be a tendency to manipulate records to mitigate potential indicators of corruption. Additionally, audits seem to play a significant role in disciplining and reducing corrupt behaviour.

We found one study that apply ML to classify social media posts for corruption. Li et al. [23] employed an Natural Language Processing (NLP) approach to gather and assess Twitter data to identify instances where users self-reported experiences of corruption, mainly within the healthcare sector. Following the analysis of tweets filtered using corruption-related keywords and using NLP techniques and manual annotation, they found 2,383 tweets. The authors clustered these tweets into actor-based topics, resulting in two main themes - police bribery and corruption in healthcare.

While existing works contribute to corruption detection, they primarily focus on identifying general corruption. In contrast, our study extends beyond this scope by considering a broader range of corruption topics, such as bribery, collusion, abuse of power, fraud, and obstruction of justice. Additionally, our pipeline incorporates fine-tuning state-of-the-art LLMs instead of relying on standard ML models. Furthermore, we explore the detection of corruption on Twitter and within the health context, offering a more comprehensive approach to addressing corruption detection.

3 Hierarchical Corruption Dialogue Taxonomy (HCDT)

As discussed in Section 2, there is a dearth of research on corruption dialogue classification using ML techniques; only one related work was identified, i.e., Li et al. [23]. Furthermore, establishing whether content pertains to corruption generally and/or a specific type of corrupt practice without a taxonomy is a significant challenge. There is little agreement on the definition of corruption let alone sub-categories of corruption [24,25]. While Li et al. [23] clustered tweets by actor-based themes, the clustering was not based on a comprehensive approach to corruption-related categories. Other categorisations proposed in the wider corruption literature are not sufficiently granular or comprehensive for use in classifying corruption dialogue. For example, Jancsics [26] categorises corruption into four types - market corruption, social bribes, corrupt organizations, and state capture. Similarly, Bussell [27] again categorises corruption into four types - legislative corruption, contracting, employment, and services. In both cases, these taxonomies are conceptual and were not applied empirically.

We organise the taxonomy on a two-level hierarchical structure. The first level is binary, i.e., *corruption* and *non-corruption.* As per Zhang et al. [28], we do not detail the non-corruption category as it is not our focus. The second level is based on sources from across all disciplines, including those from academia and practice. It contains ten specific categories and a miscellaneous or general corruption category. Table 1 presents the HCDT with definitions from Corruption Watch [29], UNODC [30] and LexisNexis [31]. In Section 4, we present terms, words, and word stems based on corrupt practices for each second level category. Again, these were sourced from commonly cited glossaries of corruption including those from Corruption Watch [29], UNODC [30], and Transparency International [32]. It should be noted that we excluded terms and practices that

Table 1: Hierarchical corruption categories with definitions

First Level	Second level	Definition
Corruption	Abuse of power/authority	"The use of one's position or authority to commit an unlawful act for the purpose of obtaining a personal advantage or an advantage for another person or entity, out of which one can derive personal gain. Abuse of power can also refer to the refusal to perform an act or function which forms part of prescribed duties." [29]
	Bribery	"The act of offering someone money, services or other valuables, in order to persuade him or her to do something in return." [29]
	Collusion/conspiracy	"Collusion is secret agreement between parties, in the public and/or private sector, to conspire to commit actions aimed to deceive or commit fraud with the objective of illicit financial gain. Conspiracy is an agreement between two or more persons to commit an offence, or which necessarily involves committing an offence." [29, 31]
	Conflict of interest	"This arises when an individual with a formal responsibility to serve the public participates in an activity that jeopardises his or her professional judgment, objectivity and independence. Often this activity primarily serves personal interests and can potentially influence the objective exercise of the individual's official duties." [29]
	Embezzlement	"When a person holding office in an institution, organisation or company dishonestly and illegally appropriates, uses or traffics the funds and goods they have been entrusted with for personal enrichment or other activities." [29]
	Extortion	"The act of utilising, either directly or indirectly, one's access to a position of power or knowledge to demand unmerited cooperation or compensation as a result of coercive threats." [29]
	Fraud	"The unlawful and intentional making of a misrepresentation which causes actual prejudice or which is potentially prejudicial to another." [29]
	Money laundering	"Any act or attempted act to disguise the source of money or assets derived from criminal activity." [29]
	Nepotism/favouritism	"Favouritism refers to the normal human inclination to prefer acquaintances, friends and family over strangers. Nepotism is a form of favouritism based on acquaintances and familiar relationships whereby someone in an official position exploits his or her power and authority to provide a job or favour to a family member or friend, even though he or she may not be qualified or deserving." [29]
	Obstruction of justice	"The use of physical force, threats or intimidation, or the promise, offering of an undue advantage to induce false testimony or to interfere in the giving of testimony or the production of evidence in a proceeding in relation to the commission of offences established in accordance with the United Nations Convention against Corruption." [30]
	General corruption	"Other types of corruption not covered in the previous categories."

were not necessarily related to a corrupt practice *per se* e.g., whistleblowers, transparency, fiduciary risk, avoiding tax, etc.

4 Data and Methods

Figure 1 illustrates our approach. Initially, we leverage the Twitter API to gather tweets about the COVID-19 vaccine discourse by employing specific keywords. Subsequently, we search for corruption- and non-corruption-related tweets and label them using both human and GPT-based coders. This annotation process leads to the creation of two distinct datasets: (1) a binary dataset with only corruption-related and non-corruption-related tweets (HCDT Level 1), and (2) a multi-class dataset consisting of 11 corruption-related topics (HCDT Level 2). Next, we fine-tune different BERT-based models using the annotated datasets, with the dual objective of first discerning whether tweets are related to corruption or not, and then identifying the specific type of corruption if present. Finally, we evaluate the performance of all models using a separate test dataset.

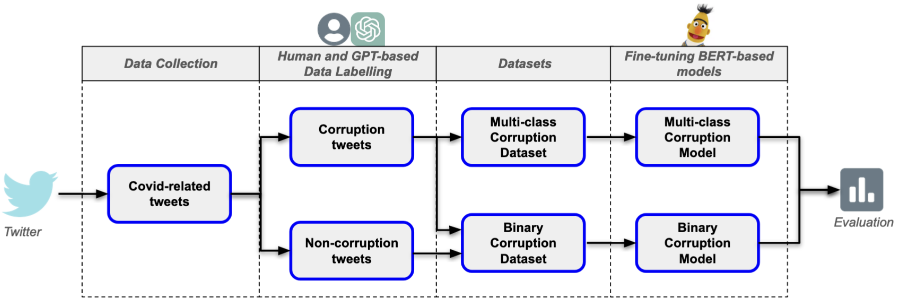

Data Collection
Human and GPT-based Data Labelling
Datasets
Fine-tuning BERT-based models
Twitter
Covid-related tweets
Corruption tweets
Non-corruption tweets
Multi-class Corruption Dataset
Binary Corruption Dataset
Multi-class Corruption Model
Binary Corruption Model
Evaluation

Fig. 1: Approach overview for automatically classifying perceived corruption in the vaccine discourse on Twitter using LLMs.

4.1 Data Collection

Data were collected on English language tweets associated with vaccines, vaccination and the COVID-19 pandemic posted on Twitter during a 12-month period between December 2020, when the COVID-19 vaccine was first released for public use in the United States, and November 2021. The following terms were used to generate the dataset: (“COVID-19” and “vaccine”) or (“coronavirus” and “vaccine”) or (“covid” and “vaccine”). Data was accessed through the Twitter enterprise API. This resulted in a dataset of 379,066,249 tweets.

Next, we generated a list of words, phrases and word stems related to corrupt practices associated with each second level category of the HCDT (see Table 2). This “bag of words" was used to filter potential tweets featuring corruption dialogue from the initial data set. Furthermore, only original tweets with at least 40 characters were considered; replies and retweets were excluded. We then pre-processed tweets by removing all URLs. After pre-processing, the dataset comprised 53,702,090 tweets.

4.2 Data Labelling

High-quality labelled data is essential for achieving good performance and generalisation when fine-tuning LLMs [33], especially in the case of text classification. However, manually labelling a large dataset can be costly, time-consuming, and prone to errors [34]. False positives are particularly prevalent when using a “bag of words" approach for filtering. To address these challenges, we employed a three-stage labelling process to identify and label tweets against both levels of HCDT.

In the first stage, two independent human coders coded tweets against HCDT Level 1 (i.e., corruption or not). If the tweets are related to corruption, the two independent coders then classified each tweet against one HCDT Level 2 category. This resulted in a dataset of 20,804 tweets.

GPT [18] has been effectively used as an annotator in natural language tasks. For example, studies suggest that the model performance trained on the GPT-3 annotated data is often comparable to or even better than trained on human-annotated data [34,35]. Therefore, in the second stage, we used GPT as an additional annotator to conduct the same exercise performed in the first stage. Specifically, we employed the following prompt using GPT-3.5-turbo: *Check whether the following tweet delimited by triple backticks is corruption-related (including topics such as: abuse of power; bribery, conspiracy, conflict of interest, embezzlement, extortion, fraud, money laundering, nepotism, and obstruction of justice) and in the coronavirus context (discussing COVID-19, vaccine, pandemic). Answer in a JSON file format according to the examples shown.*

This instruction was followed by two examples to improve the model's contextual understanding of the task. Once the GPT completed annotating the dataset, we compared the human and GPT classifications and only retained those tweets where both classifications agreed. This resulted in a dataset of 11,945 tweets.

Fine-tuning language models with balanced datasets enhances performance in downstream tasks [36, 37]. As such, in the third stage, we randomly select tweets from the entire dataset and used GPT once again to annotate a dataset of non-corruption tweets until we get the same number of tweets as the corruption dataset generated in the second stage, i.e., 11,945 tweets.

4.3 Labelled Datasets

For training and testing the proposed models, two balanced labelled datasets are required - (1) a binary corruption dataset (HCDT Level 1) and (2) a multi-class corruption dataset (HCDT Level 2). In effect, the former is a subset of the latter, and thus we discuss in this order.

Multi-class Corruption Dataset (HCDT Level 2) HCDT Level 2 comprises 11 classes - abuse of power/authority; bribery; collusion and conspiracy; conflict of interest; embezzlement; extortion; fraud; money laundering; nepotism and favouritism; obstruction of justice; and general corruption. Figure 2 shows the amount of tweets per class in the annotated dataset using the labelling strategy detailed in Section 4.2.

The *General Corruption* class is the most prevalent one, comprising 1,506 samples, while the *Obstruction of Justice* class has the least representation, with only 632 samples. The remaining classes demonstrate a more balanced distribution of samples. To ensure a fair evaluation of the model's performance across different classes, the test dataset was built with the same number of samples for each class. However, due to the imbalance in class distribution, a simple percentage-based selection wouldn't suffice. Instead, 126 samples were randomly

Table 2: Corruption terms and word stems used to filter the corruption-related tweets.

Class	Word, phrase, word stem
Abuse of Power/Authority	abuse*, misuse, exploit*, manipulate, oppress*, control, dominat*, tyranny, subjugate, usurp, maltreat*, rent seeking, rent-seeking, trade influence, trading influence, elite capture, undue advantage
Bribery	bribe*, kickback, gratuity, payola, hush money, grease, palm-greasing, backhander, inducement, incentive, graft, solicit*, scoral*, fakelaki, facilitation payment, baksheesh, gift giving, interest peddling, fcpa
Collusion/Conspiracy	collusion, conspiracy, plot*, scheme, cabal, confederacy, intrigue, secret, connivance, complicity, collud*, deep state, state capture, elite capture, bid rig*, bid rotat*
Conflict of Interest	conflict, interest, unethical, bias, influence, personal gain, self-serving, double-dealing, impropriety, trading influence, trade influence, revolving door, conflict of interest
Embezzlement	embezzl*, misappropriate, steal, pilfer, purloin, peculate, defalcate, swindle, loot, skim, siphon, divert, misappropriat*, enrich*
Extortion	extort*, blackmail, shakedown, ransom, threat, pressure, coercion, squeeze, intimidation, strong-arm, coerci*
Fraud	fraud, scam, con, swindle, deceive, hoodwink, dupe, trick,cheat, sham, counterfeit, impostor, racket, spoof, crim*, concealment, gouging, tax evasion, evade tax, evading tax, carbon cowboys, mispric*, transfer mispric*
Money Laundering	money laundering, clean, dirty money, shell company, offshore account, front company, smurfing, layering, launder*, tax haven, underground bank, secrecy jurisdiction, shell company
Nepotism/Favouritism	nepotism, favouritism, favouritism, patronage, bias, partiality, preferential treatment cronyism, crony*, old boy network, clientelis*, enrich, neopatrimon*
Obstruction of Justice	obstruction*, *justice, impede, hinder, cover-up, conceal, pervert, tamper, stonewall, spoliat*
General Corruption	corrupt*

chosen from each class. This value represents 20% of the samples from the least represented class. The remaining samples were used as training data for the multi-class corruption model, comprising 10,559 samples across the 11 HCDT Level 2 categories.

Binary Corruption Dataset (HCDT Level 1) The HCDT Level 1 dataset comprised two classes - corruption-related and non-corruption-related tweets. To build the corruption-related test set, we selected the 126 testing samples from each HDCT 2 class as detailed above, resulting in a total of 1,386 instances. In this case, all samples were simply labelled as corrupted-related. The remaining 10,559 tweets were designated as training samples. Similarly, to construct the non-corruption test dataset portion, we also randomly sampled 1,386 samples from the pool of 11,945 non-corruption-related, creating a balanced test dataset. Again, the remaining 10,559 non-corruption related were considered as training samples.

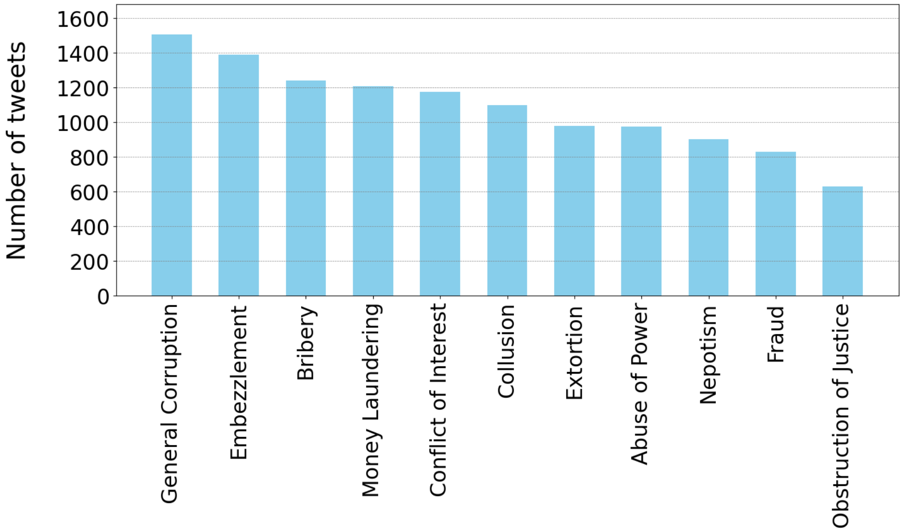


Fig. 2: Corruption-related class distribution.

4.4 Fine-tuning LLMs

After preparing the training and testing datasets, we selected the following three pre-trained encoder-based LLMs to perform fine-tuning: `BERT`[5], `RoBERTa`[6], and `BERTweet`[7]. BERT [38] is a standard LLM that achieves outstanding performance in many text classification problems [39–41]. RoBERTa is a modified version of BERT with improvements in training methodology and performance across various NLP tasks [42]. On the other hand, BERTweet [43] is a BERT-based model specifically fine-tuned on a large corpus of English tweets related to COVID-19. It is optimized for the unique features of tweets, including informal grammar, short text length, and irregular vocabulary. We decided not to use decoder-based LLMs, such as GPT [18] or Mistral [44], because they introduce challenges, such as the high costs associated with commercial APIs and computational resources and scalability issues in real-world scenarios [37].

These three models were fine-tuned for both binary and multi-class tasks discussed in this paper. The fine-tuning process was performed using the following hyperparameters: 20 epochs, 5×10^{-6} learning rate, AdamW optimizer, batch size of 8 samples, and saving the model with the highest accuracy on test data. The training and experiments were performed using a computer with Intel(R) Core(TM) i7-12700 CPU at 2.10GHz, 32 GB RAM, and Nvidia GeForce GTX 16660 SUPER.

[5] https://huggingface.co/bert-base-uncased
[6] https://huggingface.co/FacebookAI/roberta-base
[7] https://huggingface.co/vinai/bertweet-base

5 Results

Table 3 presents the LLM classification performance for the binary classification task. All models achieved over 98% for all metrics, indicating strong performance for the binary classification of corruption or non-corruption.

Table 3: Binary classification results.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
BERT	99.0260	99.0317	99.0260	99.0259
BERTweet	98.7734	98.7744	98.7734	98.7734
RoBERTa	98.6652	98.6665	98.6652	98.6652

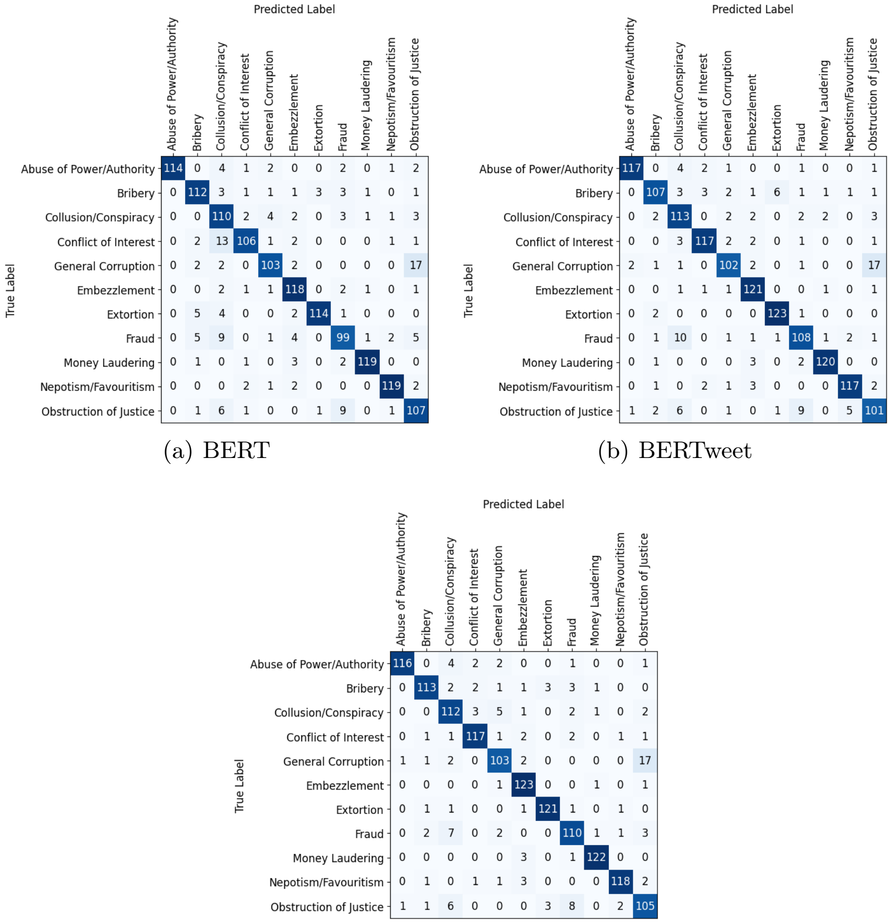
RoBERTa performed worst compared with BERT and BERTweet, with all the metrics circa 98.66%. BERTweet performed slightly better, ranking as the second-best model in our experiments. BERT demonstrated the best performance compared with the other two models, achieving 99.02% for all the metrics. All models performed to a high level across all metrics, indicating that regardless of the model, they were effective in distinguishing corruption-related content from non-corruption-related content. This is also reflective of the binary and balanced nature of the classification task.

Table 4 presents the results for the multi-class classification problem. The traditional BERT model showed the performed worst amongst the models, with an accuracy of 88.09%. BERTweet followed with slightly better results, achieving 89.89% accuracy. RoBERTa outperformed both of the other models with an accuracy of 90.90%. All the models performed best in precision, albeit slight. A model exhibiting high precision, but low recall may be preferable in situations where false positives are undesirable. In our context, it's crucial to accurately identify corruption-related content in tweets, even at the risk of overlooking some instances (resulting in lower recall). Other metrics followed a similar trend in performance across all the models.

Table 4: Multi class classification results.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
BERT	88.0952	88.7308	88.0952	88.2437
BERTweet	89.8989	90.0639	89.8989	89.9037
RoBERTa	90.9090	91.0716	90.9090	90.9350

In the multi-class classification task, model performance is lower across all metrics for all models and less consistent compared to the binary classification scenario. This outcome is expected since the classification task consists of 11 classes. Similarly, the best model for the multi-class classification task (RoBERTa) differed from the one for binary classification (BERT). Again, RoBERTa is an advancement on BERT and is better suited for the more complex task.



	Abuse of Power/Authority	Bribery	Collusion/Conspiracy	Conflict of Interest	General Corruption	Embezzlement	Extortion	Fraud	Money Laudering	Nepotism/Favouritism	Obstruction of Justice
Abuse of Power/Authority	114	0	4	1	2	0	0	2	0	1	2
Bribery	0	112	3	1	1	1	3	3	1	0	1
Collusion/Conspiracy	0	0	110	2	4	2	0	3	1	1	3
Conflict of Interest	0	2	13	106	1	2	0	0	0	1	1
General Corruption	0	2	2	0	103	2	0	0	0	0	17
Embezzlement	0	0	2	1	1	118	0	2	1	0	1
Extortion	0	5	4	0	0	2	114	1	0	0	0
Fraud	0	5	9	0	1	4	0	99	1	2	5
Money Laudering	0	1	0	1	0	3	0	2	119	0	0
Nepotism/Favouritism	0	0	0	2	1	2	0	0	0	119	2
Obstruction of Justice	0	1	6	1	0	0	1	9	0	1	107

(a) BERT

	Abuse of Power/Authority	Bribery	Collusion/Conspiracy	Conflict of Interest	General Corruption	Embezzlement	Extortion	Fraud	Money Laudering	Nepotism/Favouritism	Obstruction of Justice
Abuse of Power/Authority	117	0	4	2	1	0	0	1	0	0	1
Bribery	0	107	3	3	2	1	6	1	1	1	1
Collusion/Conspiracy	0	2	113	0	2	2	0	2	2	0	3
Conflict of Interest	0	0	3	117	2	2	0	1	0	0	1
General Corruption	2	1	1	0	102	2	0	1	0	0	17
Embezzlement	0	0	1	1	1	121	0	0	1	0	1
Extortion	0	2	0	0	0	0	123	1	0	0	0
Fraud	0	1	10	0	1	1	1	108	1	2	1
Money Laudering	0	1	0	0	0	3	0	2	120	0	0
Nepotism/Favouritism	0	1	0	2	1	3	0	0	0	117	2
Obstruction of Justice	1	2	6	0	1	0	1	9	0	5	101

(b) BERTweet

	Abuse of Power/Authority	Bribery	Collusion/Conspiracy	Conflict of Interest	General Corruption	Embezzlement	Extortion	Fraud	Money Laudering	Nepotism/Favouritism	Obstruction of Justice
Abuse of Power/Authority	116	0	4	2	2	0	0	1	0	0	1
Bribery	0	113	2	2	1	1	3	3	1	0	0
Collusion/Conspiracy	0	0	112	3	5	1	0	2	1	0	2
Conflict of Interest	0	1	1	117	1	2	0	2	0	1	1
General Corruption	1	1	2	0	103	2	0	0	0	0	17
Embezzlement	0	0	0	0	1	123	0	0	1	0	1
Extortion	0	1	1	0	0	1	121	1	0	1	0
Fraud	0	2	7	0	2	0	0	110	1	1	3
Money Laudering	0	0	0	0	0	3	0	1	122	0	0
Nepotism/Favouritism	0	1	0	1	1	3	0	0	0	118	2
Obstruction of Justice	1	1	6	0	0	0	3	8	0	2	105

(c) RoBERTa

Fig. 3: Confusion matrices for the different BERT-based models.

Figure 3 shows the confusion matrices for the multi-class classification task. The diagonals represent correct predictions. The models classified most tweets in the test dataset accurately. RoBERTa achieved higher values in the matrices diagonals for most classes. The other models outperformed RoBERTa in some classes. For instance, BERT outperformed other models for *nepotism and*

Table 5: Examples of models prediction (we replaced the original usernames by @user in order to keep the confidentiality).

Tweet	Label	BERT	BERTweet	RoBERTa
The FDA Cover-up that Led to the Approval of the Pfizer Vaccine	Corruption	Obstr. Just.	Obstr. Just.	Obstr. Just.
@user It's blatantly obvious anyone pushing vaccines this hard has had their palms greased	Bribery	Fraud	Embez.	Bribery
@user So state looted the covid funds thinking vaccines were for free	Embez.	Embez.	Embez.	Embez.

favouritism and *obstruction of justice*. On the other hand, BERTweet outperformed other models for *collusion and conspiracy* and *extortion*.

Certain classes had a higher frequency of misclassifications. For instance, some tweets labelled as *conflict of interest* were misclassified as *collusion and conspiracy*, with BERT, BERTweet and RoBERTa misclassifying 13, 3, and 1 tweets, respectively. Similarly, *fraud* was mistaken with *collusion and conspiracy* by BERT (9 times), BERTweet (10 times), and RoBERTa (7 times). The highest number of incorrect classifications occurred when confusing *general corruption* with *obstruction of justice*, representing 17 misclassifications across all models. Due to the relatively small number of instances, a manual review was undertaken. The primary cause of the misclassification would seem to related to a combination of overlapping vocabulary and semantic ambiguity. For example, terms and phrases in the context of obstruction of justice and general corruption were common in both, e.g., cover-up and tamper. Furthermore, corruption-related language can be vague and coded, e.g., making use of euphemisms or indirect language, thus making it hard for a model to understand the intent. This issue may be resolved using a larger and more diverse training dataset, and multi-label classification rather than merely multi-class classification.

Table 5 shows some instances of corruption-related tweets with the classification of each model. The first tweet implies a corrupt behaviour within the FDA approval process. Therefore, we labelled it as *general corruption* since it suggests that information was concealed or the approval process was manipulated. All models misclassified it as *obstruction of justice*, which would be understandable if the tweet's focus was on deliberate evidence concealment. We labelled the second tweet as *bribery* class since the phrase "palms greased" suggests the idea of someone receiving financial incentives or benefits in exchange for promoting vaccines. In this case, only the RoBERTa model correctly classified this tweet. Finally, we labelled the third tweet as *embezzlement* since the mention of "state looted the covid funds" implies that the funds allocated for COVID-related purposes were diverted for other uses, which aligns with the concept of embezzlement. All models correctly predicted this tweet.

6 Limitations and future work

This study is not without limitations, which in themselves provide future avenues for research. Firstly, the study was limited to one social media platform, Twitter, one language, English, and one health and vaccination context, COVID-19 and vaccination. Since the acquisition of Twitter, the owners of X have introduced new cost-based API access for researchers. This is a significant economic barrier to similar research moving forward and may encourage less ethical access to data through scraping without permission. Also, there exists significant potential for undertaking similar work on other social media platforms and specifically those with research APIs e.g., TikTok. Similarly, while English is widely spoken worldwide, it is not representative of many of the countries and regions most severely impacted by COVID-19. While the authors plan to localise the HCDT for Portuguese, Spanish, and Italian and develop associated annotated datasets and models, similar work is required for other major languages.

Secondly, the HCDT is limited to two levels and while a comprehensive set of indicative practices per second level category was identified to create a 'bag of words', the taxonomy could be extended further. For example, other corruption typologies could be integrated, and a third class could be introduced at the first level, i.e., anti-corruption, which would facilitate more nuanced analysis.

Thirdly, we identified some misclassification issues, which we believe are due to overlapping vocabulary and semantic ambiguity. While performance on binary and multi-class classification tasks was relatively high, even better performance may be achieved by increasing the size and diversity of the annotated datasets, applying multi-label classification or exploring ensemble solutions, which we plan to address in future works.

7 Conclusion

Perceived or real corruption can negatively impact trust in health services and associated outcomes. Such perceptions can be amplified and propagated on social media and interfere with public health communication. In the case of vaccination, this can result in lower levels of immunisation which can lead to illness and death. In this paper, we addressed the detection and classification of perceived corruption in the COVID-19 vaccine discourse on social media.

First, we proposed a hierarchical corruption dialogue taxonomy (HCDT), a two-level hierarchical taxonomy of corruption dialogue that can be used for categorising content by type of corruption. Second, to support future research in the identification and classification of corruption dialogue in the COVID-19 discourse on Twitter, we developed a labelled dataset for training and testing classification models. Third, we evaluated three different pre-trained BERT-based architectures (BERT, RoBERTa, and BERTweet) for (i) classifying tweets as corruption-related or non-corruption-related, and (ii) classifying tweets as one of 11 types of corruption according to the HCDT. For the binary classification, all the models obtained similar performance, with metrics around 98% and 99%;

BERT marginally outperformed the other models. For the multi-class classification task, RoBERTa outperformed other models, with all the metrics around 91%.

Acknowledgment

This work was partially funded by the Irish Institute of Digital Business (IIDB) at Dublin City University, Ireland.

References

1. World Health Organisation, "WHO COVID-19 dashboard," 2024.
2. World Health Organisation (WHO), "Global covid-19 vaccination strategy in a changing world july 2022 update," 2022.
3. P. F. Burke, D. Masters, and G. Massey, "Enablers and barriers to covid-19 vaccine uptake: An international study of perceptions and intentions," *Vaccine*, vol. 39, no. 36, pp. 5116–5128, 2021.
4. K. M. Jia, W. P. Hanage, M. Lipsitch, A. G. Johnson, A. B. Amin, A. R. Ali, H. M. Scobie, and D. L. Swerdlow, "Estimated preventable covid-19-associated deaths due to non-vaccination in the united states," *European Journal of Epidemiology*, vol. 38, no. 11, pp. 1125–1128, 2023.
5. M. Zhong, M. Kshirsagar, R. Johnston, R. Dodhia, T. Glazer, A. Kim, D. Michael, S. Nair-Desai, T. C. Tsai, S. Friedhoff *et al.*, "Estimating vaccine-preventable covid-19 deaths under counterfactual vaccination scenarios in the united states," *medRxiv*, pp. 2022–05, 2022.
6. P. Eigen, "Measuring and combating corruption," *The Journal of Policy Reform*, vol. 5, no. 4, pp. 187–201, 2002.
7. Transparency International, "What is corruption?" 2024.
8. N. Naher, R. Hoque, M. S. Hassan, D. Balabanova, A. M. Adams, and S. M. Ahmed, "The influence of corruption and governance in the delivery of frontline health care services in the public sector: a scoping review of current and future prospects in low and middle-income countries of south and south-east asia," *BMC public health*, vol. 20, pp. 1–16, 2020.
9. M. R. Farzanegan and H. P. Hofmann, "Effect of public corruption on the covid-19 immunization progress," *Scientific reports*, vol. 11, no. 1, p. 23423, 2021.
10. J. C. Kohler, "Covid-19 vaccines and corruption risks: preventing corruption in the manufacture, allocation and distribution of vaccines," 2020.
11. R. K. Goel, M. A. Nelson, and V. Y. Goel, "Covid-19 vaccine rollout—scale and speed carry different implications for corruption," *Journal of Policy Modeling*, vol. 43, no. 3, pp. 503–520, 2021.
12. A. Spreco, T. Schön, and T. Timpka, "Corruption should be taken into account when considering covid-19 vaccine allocation," *Proceedings of the National Academy of Sciences*, vol. 119, no. 19, p. e2122664119, 2022.
13. A. Egli, P. Rosati, T. Lynn, and G. Sinclair, "Bad robot: A preliminary exploration of the prevalence of automated software programmes and social bots in the covid-19# antivaxx discourse on twitter," in *Proceedings of the The International Conference on Digital Society, Nice, France*, 2021, pp. 18–22.

14. D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, “Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate,” *American journal of public health*, vol. 108, no. 10, pp. 1378–1384, 2018.
15. F. Rodrigues, N. Ziade, K. Jatuworapruk, C. V. Caballero-Uribe, T. Khursheed, and L. Gupta, “The impact of social media on vaccination: A narrative review,” *Journal of Korean Medical Science*, vol. 38, no. 40, 2023.
16. S. L. Wilson and C. Wiysonge, “Social media and vaccine hesitancy,” *BMJ global health*, vol. 5, no. 10, p. e004206, 2020.
17. I. Skafle, A. Nordahl-Hansen, D. S. Quintana, R. Wynn, and E. Gabarron, “Misinformation about covid-19 vaccines on social media: rapid review,” *Journal of medical Internet research*, vol. 24, no. 8, p. e37367, 2022.
18. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
19. M. S. M. Lima and D. Delen, “Predicting and explaining corruption across countries: A machine learning approach,” *Government information quarterly*, vol. 37, no. 1, p. 101407, 2020.
20. K. Rabuzin and N. Modrusan, “Prediction of public procurement corruption indices using machine learning methods,” in *KMIS*, 2019, pp. 333–340.
21. E. Denisova-Schmidt, M. Huber, E. Leontyeva, and A. Solovyeva, “Combining experimental evidence with machine learning to assess anti-corruption educational campaigns among russian university students,” *Empirical Economics*, vol. 60, pp. 1661–1684, 2021.
22. E. Ash, S. Galletta, and T. Giommoni, “A machine learning approach to analyzing corruption in local public finances,” *Center for Law & Economics Working Paper Series*, vol. 6, 2020.
23. J. Li, W.-H. Chen, Q. Xu, N. Shah, J. C. Kohler, and T. K. Mackey, “Detection of self-reported experiences with corruption on twitter using unsupervised machine learning,” *Social Sciences & Humanities Open*, vol. 2, no. 1, p. 100060, 2020.
24. A. Graycar, “Corruption: Classification and analysis,” *Policy and Society*, vol. 34, no. 2, pp. 87–96, 2015.
25. J. Rose, “The meaning of corruption: Testing the coherence and adequacy of corruption definitions,” *Public Integrity*, vol. 20, no. 3, pp. 220–233, 2018.
26. D. Jancsics, “Corruption as resource transfer: An interdisciplinary synthesis,” *Public Administration Review*, vol. 79, no. 4, pp. 523–537, 2019.
27. J. Bussell, “Typologies of corruption: A pragmatic approach,” in *Greed, corruption, and the modern state.* Edward Elgar Publishing, 2015, pp. 21–45.
28. Y. Zhang, P. Ren, and M. de Rijke, “A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses,” *Journal of the Association for Information Science and Technology*, vol. 72, no. 12, pp. 1477–1497, 2021.
29. Corruption Watch, “Glossary of corruption-related terms,” 2022.
30. UNODC, “Glossary of corruption-related terms,” 2019.
31. LexisNexis, “Glossary,” 2024.
32. Transparency International, “Corruption A-Z,” 2024.
33. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
34. S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, “Want to reduce labeling cost? gpt-3 can help,” *arXiv preprint arXiv:2108.13487*, 2021.

35. B. Ding, C. Qin, L. Liu, Y. K. Chia, S. Joty, B. Li, and L. Bing, “Is gpt-3 a good data annotator?” *arXiv preprint arXiv:2212.10450*, 2022.
36. R. Zevallos, M. Farrús, and N. Bel, “Frequency balanced datasets lead to better language models,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 7859–7872.
37. V. G. dos Santos, G. L. Santos, T. Lynn, and B. Benatallah, “Identifying citizen-related issues from social media using llm-based data augmentation,” in *International Conference on Advanced Information Systems Engineering*. Springer, 2024, pp. 531–546.
38. 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.
39. J.-S. Lee and J. Hsiang, “Patent classification by fine-tuning bert language model,” *World Patent Information*, vol. 61, p. 101965, 2020.
40. G. L. Santos, V. G. dos Santos, C. Kearns, G. Sinclair, J. Black, M. Doidge, T. Fletcher, D. Kilvington, P. T. Endo, K. Liston *et al.*, “Kicking prejudice: Large language models for racism classification in soccer discourse on social media,” in *International Conference on Advanced Information Systems Engineering*. Springer, 2024, pp. 547–562.
41. S. Gupta, S. Bolden, J. Kachhadia, A. Korsunska, and J. Stromer-Galley, “Polibert: Classifying political social media messages with bert,” in *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*, 2020.
42. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
43. D. Q. Nguyen, T. Vu, and A. T. Nguyen, “Bertweet: A pre-trained language model for english tweets,” *arXiv preprint arXiv:2005.10200*, 2020.
44. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, “Mistral 7b,” *arXiv preprint arXiv:2310.06825*, 2023.