Classifying Human-Generated and AI-Generated Election Claims in Social Media

Alphaeus Dmonte[1], Marcos Zampieri[1], Kevin Lybarger[1], Massimiliano Albanese[1] and Genya Coulter[2]

[1]*Department of Information Sciences and Technology, George Mason University, USA*
[2]*OSET Institute*
{*admonte, mzampier, klybarge, malbanes*}*@gmu.edu, genya@osetinstitute.org*

Abstract: Politics is one of the most prevalent topics discussed on social media platforms, particularly during major election cycles, where users engage in conversations about candidates and electoral processes. Malicious actors may use this opportunity to disseminate misinformation to undermine trust in the electoral process. The emergence of Large Language Models (LLMs) exacerbates this issue by enabling malicious actors to generate misinformation at an unprecedented scale. Artificial intelligence (AI)-generated content is often indistinguishable from authentic user content, raising concerns about the integrity of information on social networks. In this paper, we present a novel taxonomy for characterizing election-related claims. This taxonomy provides an instrument for analyzing election-related claims, with granular categories related to jurisdiction, equipment, processes, and the nature of claims. We introduce ElectAI, a novel benchmark dataset that consists of 9,900 tweets, each labeled as human- or AI-generated. For AI-generated tweets, the specific LLM variant that produced them is specified. We annotated a subset of 1,550 tweets using the proposed taxonomy to capture the characteristics of election-related claims. We explored the capabilities of LLMs in extracting the taxonomy attributes and trained various machine learning models using ElectAI to distinguish between human- and AI-generated posts and identify the specific LLM variant.

1 Introduction

The widespread use of social media has fundamentally changed political discourse with direct impact on elections. Social media enables candidates and political entities to directly communicate with the electorate through platforms, such as X (formerly Twitter), profoundly shaping engagement strategies. This direct communication channel facilitates the rapid propagation of claims about election processes, including claims of fraud, by politicians and their supporters. Claims of fraud that are intentionally false or not fully supported by credible evidence can negatively impact election processes and compromise the integrity of elections. As a result, elections in many countries, such as the United States (US), have become more contentious in recent years[1].

The emergence of Large Language Models

---
[1]https://www.pewresearch.org/short-reads/2020/03/04/far-more-americans-see-very-strong-partisan-conflicts-now-than-in-the-last-two-presidential-election-years/

(LLMs) has revolutionized content generation across various domains. LLMs are capable of generating fluent and syntactically correct text that is virtually indistinguishable from human-generated text. LLMs often generate factually incorrect statements and, in some cases, may hallucinate (Huang et al., 2023). For example, Galactica (Taylor et al., 2022), an LLM developed by Meta to summarize research papers and generate Wikipedia articles, was found to generate biased and incorrect content (Heaven, 2022). ChatGPT was also found to produce biased and inaccurate outputs (Wach et al., 2023). Furthermore, malicious actors can leverage the power of LLMs to intentionally produce uninformed claims and misinformation to spread them on social media (Zhou et al., 2023).

The convergence of advanced LLMs and the widespread use of social media creates the perfect conditions for the spread of election misinformation. In this paper, we introduce a comprehensive taxonomy to further our understanding of human- and AI-generated election claims in social media. The taxonomy features several dimensions that often char-

acterize election claims, such as jurisdiction, election infrastructure, and the nature of claims. We use the taxonomy to annotate ElectAI, the first benchmark dataset for election claim understanding. ElectAI contains a mix of 9,900 human- and AI-generated tweets in English.

We present several experiments on claim characterization and authorship identification using ElectAI with the objective to explore the following research questions:

- **RQ1:** How well do state-of-the-art LLMs understand the various attributes of election claims?

- **RQ2:** How do humans and machines compare in their ability to distinguish between human- and AI-generated election claims in social media?

Our main contributions are as follows:

1. We propose a novel taxonomy to characterize and understand election-related claims and the attributes associated with these claims.

2. We present ElectAI, the first benchmark dataset for election claims. The dataset comprises both human- and AI-generated tweets annotated with the proposed taxonomy. We make ElectAI freely available to the research community.[2]

3. We present an evaluation of four LLMs (Llama-2, Mistral, Falcon, and Flan-T5) with respect to claim characterization on the ElectAI dataset.

4. We evaluate the capability of multiple models and humans to discriminate between human- and AI-generated posts in the ElectAI dataset.

The remainder of the paper is organized as follows. Section 2 discusses the importance of fair and secure elections, motivating our choice of US elections as a case study, whereas Section 3 discusses related work. Then, Section 4 introduces the proposed taxonomy and Section 5 presents the ElectAI dataset. Next, Sections 6 and 7 present our evaluation of claim understanding and authorship attribution, respectively. Finally, Section 8 provides some concluding remarks and future research directions.

2   Background and Motivation

Fair and secure elections are the backbone of democracy, and the spread of misinformation on social media can undermine their integrity. In 2017, former US Secretary of the Department of Homeland Security (DHS), Jeh Johnson, declared that *"election infrastructure in this country should be designated as a*

subsector of the existing Government Facilities critical infrastructure sector. Given the vital role elections play in this country, certain systems and assets of election infrastructure meet the definition of critical infrastructure, in fact and in law."[3] As a result, US election infrastructure is considered critical infrastructure, as much as energy, transportation, or other systems critical to national security. Threats to election security can erode voter trust, delegitimize elections, and weaken democratic institutions. The spread of misinformation, disinformation, and malinformation (MDM) is a growing threat to fair and secure elections. MDM campaigns can target individual political candidates, local or state administrative entities, or even manufacturers of voting equipment. MDM campaigns erode trust in election processes and can also have significant financial consequences.

While the proposed approach has been demonstrated in the context of US elections, it can be easily generalized to create and study election-related datasets in any language or geopolitical context. American elections provide an interesting case study as they are inherently different from almost any type of election in other countries. With nearly 10,000 unique election jurisdictions, no other country has the level of decentralization that US elections do, and no other nation cedes as much control of federal elections to state and local bodies. As of 2020, over 260 million Americans were registered to vote. In jurisdictions such as Los Angeles County, California, voters will vote for more contests and candidates in one general election than the average voter overseas might vote for in their lifetime.

As elections are conducted at the local level, there is great variability in the infrastructure and technical expertise of counties and towns within states. While the decentralization of elections avoids single points of failure, it requires each jurisdiction to manage and secure an extremely complex information technology system that includes voter registration databases, electronic poll books, vote-capture devices, optical scanners, vote tallying devices, election night reporting systems, and network technologies to securely transfer data. The complexity and diversity of this infrastructure further compound the challenges in assessing election-related claims. Thus, the US is the ideal test case for ElectAI.

Another dimension contributing to the uniqueness of US elections is the emphasis on freedom of expression, allowing anyone to say virtually anything in any public forum. The United States of America is the

---

[2]https://languagetechnologylab.github.io/ElectAI/

---

[3]Department of Homeland Security (2017). Statement by Secretary Jeh Johnson on the Designation of Election Infrastructure as a Critical Infrastructure Subsector.

world's longest lasting representative democracy, and has served as a model to countless other nations inspired by the uniquely American ideals of independence, self-determination, and free expression. For voters who live in the United States, there are very few restrictions on political speech by design. The First Amendment to the U.S. Constitution enshrines strong historic protections of freedom of expression, whether it is speech by individuals or by the press, free practice of religion, the right of assembly, as well as the right to petition the government for a redress of grievances.

At the nexus of these widely admired but difficult to sustain ideals, is the social media platform now known as X, formerly Twitter, nimbly intersecting First Amendment rights and protections. While not as large as competing platforms such as TikTok or Facebook, it is fair to postulate that the cumulative impact of documenting current geopolitical events in real time with comparatively few barriers to entry has permanently enshrined X into the American lexicon. Elections, and the manner in which information about them is disseminated via social media channels, have been fundamentally transformed by the X platform. More Americans get their election information via X than they do from previously trusted sources such as local newspapers or from political parties. X offers an unprecedented level of instantaneous access to a voter's local election officials, as well as direct communication with state and federal authorities. Failing to quantify the impact of AI-generated election misinformation on social media users now will set media literacy, public confidence in elections, and potentially, the stability of democratic processes themselves in imminent danger. By design, ElectAI allows for a narrowing or broadening of scope, without limiting its application to just one field of research.

3 Related Work

There have been several studies addressing automatic verification of claims, rumor detection, and several other tasks related to the detection and mitigation of misinformation. Chen *et al.* (Chen et al., 2022) proposed a 3-stage pipeline for claim verification that (i) gets the encodings for the claims, (ii) retrieves relevant documents that main contain a support claim, and (iii) finds appropriate evidence in them. Yang *et al.* (Yang et al., 2022) propose a Multi-Instance Learning (MIL) approach for rumor and stance detection. Barbera *et al.* (Barbera, 2018) explored misinformation spread and the factors that contribute to it in the context of the 2016 US Presidential elections.

Recent developments in fact verification systems have significantly improved the accuracy and efficiency of automatic claim verification in various settings. Hu *et al.* (Hu et al., 2023) proposed a fact-verification approach based on training an evidence retriever and a claim verifier with the retrieved evidence. The approach considers the faithfulness (the model's decision-making process) and plausibility (convincing to humans) of the retrieved evidence. While the great majority of studies on fact verification are on English texts, an approach for Arabic claim verification in social media was proposed by (Sheikh Ali et al., 2023). In the proposed pipeline, they first identify a claim presented in a post and then evaluate the trustworthiness of the claim. Multimodal fact-checking has also been explored in a few studies such as those by Yao *et al.* (Yao et al., 2023) and by Sun *et al.* (Sun et al., 2023), who introduced one of the datasets described next.

There have been various datasets developed for claim understanding and verification. One of the most widely-used datasets is FEVER (Thorne et al., 2018) developed for fact extraction and verification. Another dataset developed by Sun *et al.* (Sun et al., 2023) targets multi-modal misinformation in the medical domain. The dataset contains multi-modal instances from news and tweets as well as instances generated by LLMs. A recent study by (Zhou et al., 2023) targets COVID-19 misinformation. The authors generate COVID-19-related misinformation using GPT-3 and they analyze the linguistic features of human- and AI-generated texts. The analysis indicates significant differences between the human and AI-generated text concerning the writing style, emotional tone, use of specific keywords, use of informal language, and several other features.

To the best of our knowledge, no other dataset has been created to address election claim understanding thus far. Furthermore, only a few datasets containing human- and AI-generated content have been curated for claim understanding. This motivates us to introduce ElectAI and make it freely available to the research community.

4 Taxonomy

The proposed taxonomy, shown in Figure 1, was developed by identifying the most common set of attributes characterizing social media discourse about elections, and it was validated by subject matter experts in election administration. Two of the authors are regular attendees of an annual US-based election-focused conference bringing together academia, gov-
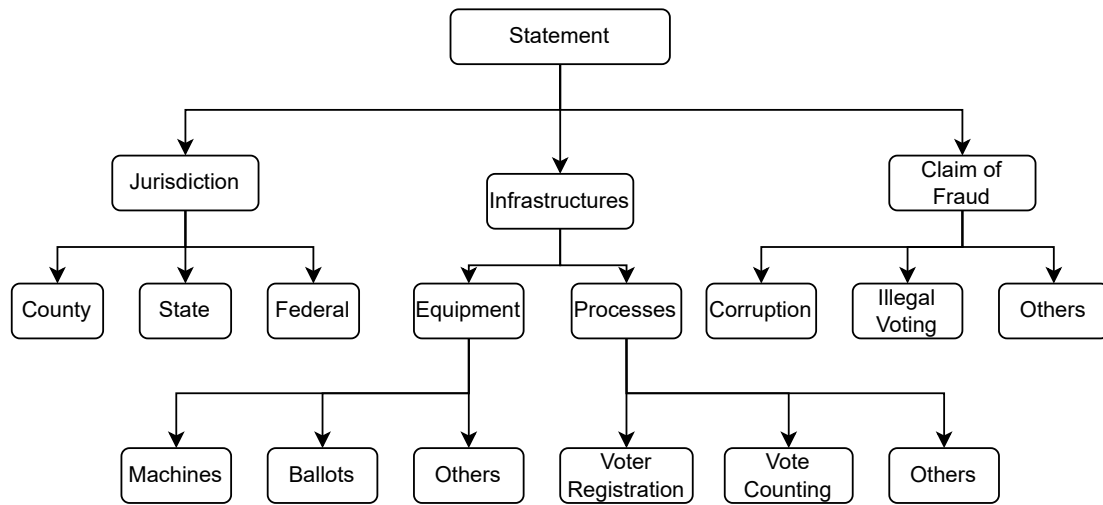
Figure 1: The election claim taxonomy.

ernment, industry, and non-profit sector, and one author is a social media expert affiliated with a nonpartisan election technology research organization working to increase confidence in elections. Details are omitted to preserve the anonymity of the submission.

The proposed taxonomy intentionally employs unambiguous naming conventions, with classification structures that can be commonly understood by a diverse group of academic researchers and election stakeholders, as well as applied to elections globally. While identical terms may not be employed worldwide, the foundational concepts of election jurisdiction, equipment, processes, and claims are commonly used by the election community globally and are highly relevant topics to the rapidly evolving field of AI and social media research. The proposed taxonomy will assist in the creation of a working set of comparative standards for future AI research endeavors while facilitating the efficient collection of quantifiable, high-quality training data for future machine learning initiatives. The race to develop data-driven methods for protecting election administrators and the voting public from the increasingly dangerous aftereffects of social media disinformation (particularly MDM generated by AI models) will be critical during the 2024 Presidential Election, when unprecedented amounts of malicious election-related social media content pose the risk of derailing American elections via ideological capture.

4.1 Jurisdiction

The Jurisdiction attribute refers to the government entity conducting the election and includes three subcategories:

- **County**: Refers to a specific county, such as Fairfax County, Los Angeles County, or Monroe County.
- **State**: Refers to a specific state, such as New York (NY), Virginia (VA), or California (CA).
- **Federal**: Refers to federal elections, namely presidential elections.

4.2 Infrastructure

The Infrastructure attribute encompasses an array of components employed to carry out elections and includes the Equipment and Processes subcategories.
Equipment. Equipment refers to any system that is used to cast or count votes and includes three subcategories:

- **Machines**: Refers to electronic voting equipment, such as direct recording electronic systems (DREs), ballot marking devices (BMDs), scanners, and electronic poll books (e-poll books).
- **Ballots**: Refers to paper ballots, including mail-in ballots.
- **Other**: Refers to any other type of voting equipment not captured by the Machines or Ballots subcategories.

Processes. Processes refers to activity conducted during the administration of elections and includes three subcategories:

- **Voter Registration**: Refers to the process of registering eligible voters in a centrally managed voter registration database.
- **Vote Counting**: Refers to the process of tallying and tabulating votes cast in an election. This

includes verifying the authenticity of ballots and systematically counting the votes to determine the outcome of the election.

- **Other**: Refers to any other election process not captured by the Voter Registration or Vote Counting subcategories.

4.3 Claim of Fraud

The Claim of Fraud attribute refers to claims of election-related fraud in a media post and includes three subcategories:

- **Corruption**: Refers to claims of corruption (e.g., a candidate offering money or favors in exchange for votes).

- **Illegal Voting**: Refers to claims of illegal voting (e.g., votes cast by illegal aliens or individuals voting multiple times).

- **Other**: Refers to any other claims of fraud not captured by the Corruption or Illegal Voting subcategories (e.g., ballot tampering or coercion tactics).

5 The ElectAI Dataset

5.1 Data Collection and Generation

The ElectAI dataset comprises a total of 9,900 tweets that were either human- or AI-generated. All tweets include a label indicating whether they are created by a human or AI, and all AI-generated tweets include a label indicating the LLM variant that was used to create them. A subset of 1,550 tweets were annotated using the proposed taxonomy, as reported in Table 1, including 850 human-generated and 700 AI-generated tweets.

	Human	AI	Total
First Annotation Round	600	0	600
Second Annotation Round	250	700	950
Total	850	700	1,550

Table 1: Summary of annotated posts

To populate the human-generated part of the dataset, we initially sampled the existing Voter-Fraud2020 dataset (Abilov et al., 2021), a multimodal dataset consisting of posts about potential events during the 2020 US presidential elections. Since the VoterFraud2020 dataset also includes tweets that are general statements, we use specific keywords

like jurisdiction, equipment, process, device, scanner, counting, registration, corruption, etc. to identify and sample 850 tweets relevant to our task. We use the proposed taxonomy to create narrative frames, as described in recent work on COVID-19 (Zhou et al., 2023), to generate 700 synthetic tweets using LLMs. These tweets were generated after the first round of annotations. The taxonomy attributes were used to generate tweets that spanned a diverse range of election-related topics and mimicked human-generated tweets. We use the following LLMs to generate the tweets.

Falcon-7B-Instruct (Penedo et al., 2023) henceforth Falcon, is a decoder-only model fine-tuned with instruct and chat datasets. This model was adapted from the GPT-3 (Brown et al., 2020) model, with differences in the positional embeddings, attention, and decoder-block components. The base model Falcon-7B, on which this model was fine-tuned, outperforms other open-source LLM models like MPT-7B and RedPajama, among others. The limitation of this model is that it was mostly trained on English data, and hence, it does not perform well in other languages.

Llama-2-7B-Chat (Touvron et al., 2023) henceforth Llama-2, is an auto-regressive language model with optimized transformer architecture. This model was optimized for dialogue use cases. The model was trained using publicly available online data. The model outperforms most other open-source chat models and has a performance similar to models like Chat-GPT. This model, however, works best only for English.

Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) henceforth Mistral, is a transformer-based large language model which uses two architectures coupled together, a grouped-query attention along with sliding-window attention. The model uses a byte-fallback tokenizer. The model outperforms several open-source LLMs including Llama-2-13B on various NLP tasks.

The following prompt was used to generate the tweets.

> *Question: Write {**# of tweets**} unique tweets using informal language and without the use of opinion statements, declarative statements, and call-to-action statements, about {**claim of fraud**} caused by {**infrastructure**} in {**state**}. You may choose the actual county in the state mentioned. You may include actual websites, people's names, and usernames.*
> *Answer:*

The prompt includes the *claim of fraud*, *infrastructure*, and *state* attributes of the taxonomy. The labels from the taxonomy were substituted for prompt attributes, such as equipment and processes instead of infrastructure, corruption, and illegal voting instead of the claim of fraud, and mentioning states like Arizona, Virginia, etc. We only use the state-level jurisdictions as this better mimics the human tweets. However, the prompt includes instructions to include the county jurisdiction for some tweets. The prompt also includes specific instructions regarding the number of tweets to be generated. We use this dataset consisting of 1,550 human- and AI-generated tweets as an evaluation dataset for the claim understanding task, where we extract the claim-related attributes from the tweets.

In addition to exploring the extraction of claim-related attributes from human- and AI-generated tweets, we explored the ability of humans and machine learning models to distinguish between human- and AI-generated tweets. To facilitate this exploration, we generate additional tweets with Llama-2, Falcon, and Mistral using the prompt described above. For the human-generated portion, we sample the VoterFraud2020 dataset. Our training dataset for this authorship attribution task consists of 8,000 tweets and each tweet is labeled with *human*, *llama*, *falcon*, or *mistral* label, depending on the model used to generate the tweet. Table 2 shows example instances from the training dataset.

The claim characterization dataset is reused as a test dataset for the authorship attribution task. However, for this task, we re-annotate each tweet of the test dataset with a single label, depending on the model used to generate the tweet. We generate an additional 350 tweets using the Mistral model to be included in the test dataset.[4]

---

[4]These tweets are not annotated using the taxonomy described above as the Mistral model was not publicly available at the time of annotations. However, we included these these tweets in the authorship attribution task to have instances with all the labels, we include these these tweets.

Tweet	Label
@RealNews: Why are people trying to vote twice in Arkansas? Isn't that illegal? #VoterFraud #BaldFacedLie	human
Breaking news out of Guilford County! It looks like someone has been tampering with ballots! Stay vigilant and follow @integritywatchdog for updates. #GuilfordCounty #ElectionCorruption #BallotTampering	llama
Voting rights in #Iowa are in jeopardy! Anyone with information about voter registration irregularities or suspicious activity should contact @Iowa_Election immediately - @InformedIowan	falcon
You know what's not cool? Voter registration irregularities in Kent County, Rhode Island. Hopefully the Board of Elections can sort this out before it's too late! #CleanVoterRolls	mistral

Table 2: Example tweets from the training dataset.

5.2 Annotation

We first developed a preliminary taxonomy that captured information related to jurisdiction, infrastructure, processes, and claims of fraud, similar to the final taxonomy presented in Figure 1. The preliminary taxonomy had more granular attributes (36 leaf nodes) than the taxonomy presented in Figure 1 (12 leaf nodes). Using this preliminary taxonomy, a team of undergraduate and graduate students annotated 600 of the 850 human-generated tweets sampled from the VoterFraud2020 dataset. Each student annotated a sample of 50 tweets. Each tweet was annotated by up to three annotators, and a majority vote was used to adjudicate discrepancies and create the gold reference standard. We used this initial annotation effort to refine and consolidate the preliminary taxonomy to create the taxonomy shown in Figure 1.

Table 3 presents the inter-annotator agreement (IAA) for the taxonomy attributes, as well as the overall score.

We calculate IAA for the doubly annotated samples using F1-score, by holding one annotator as the reference and the other annotator as the prediction. We use F1 to provide a metric that can be compared with our classification results. We then used the revised taxonomy and guidelines to annotate an additional 250 human-generated tweets along with the 700 AI-generated ones, for a total of 950 tweets in round 2. Each tweet was annotated by up to three annotators. The data were annotated by undergraduate and graduate students in information technology and computer science. We provided detailed guidelines crafted based on the taxonomy, and the tweets were assigned a label using a majority vote.

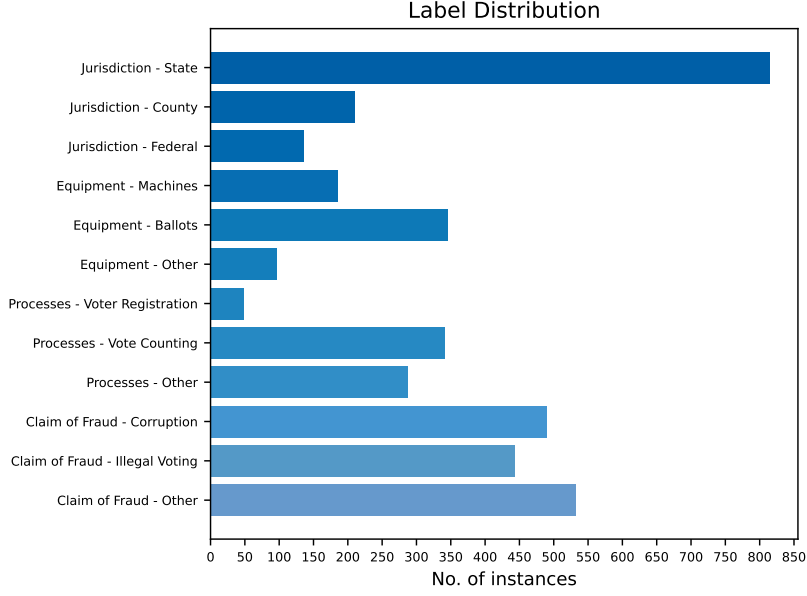As can be seen in Table 3, there is a substantial to

Figure 2: Statistics of the annotations. These statistics are based on the two rounds of annotation.

a high agreement for labels Jurisdiction -State, Jurisdiction - County, Equipment - Machines, Equipment - Ballots, Claim of Fraud - Corruption, and Claim of Fraud - Illegal Voting. However, the labels Jurisdiction - Federal, Equipment-Others, Processes - Voter Registration, Processes - Vote Counting, Processes - Other, and Claim of Fraud - Other, had fair to slight agreement. An overall F1-score of 0.59 indicates a moderate agreement between the annotators over all the labels. This level of agreement can be attributed to a low agreement for some labels.

Label	F1
Jurisdiction - State	0.86
Jurisdiction - County	0.66
Jurisdiction - Federal	0.28
Equipment - Machines	0.79
Equipment - Ballots	0.70
Equipment - Other	0.04
Processes - Voter Registration	0.16
Processes - Vote Counting	0.53
Processes - Other	0.21
Claim of Fraud - Corruption	0.68
Claim of Fraud - Illegal Voting	0.65
Claim of Fraud - Other	0.45
Overall	**0.59**

Table 3: IAA, measured as F1. Classes included in the claim understanding experiments are presented in bold.

5.3 Data Statistics

The 600 tweets from the first annotation effort were combined with the 950 tweets annotated using the refined taxonomy, to get a dataset with 1,550 annotated tweets, that we use for the claim understanding task. To maintain consistency between the annotations, we map the annotations from the first annotation effort to the refined taxonomy. Figure 2 shows the statistics of the dataset for each label. As seen, most of the tweets had a mention of Jurisdiction-State. Whereas, not many tweets mention Processes-Voter Registration, Equipment-Others, and Jurisdiction-Federal. The mention of the various claims of fraud is almost evenly distributed between the tweets. We observed that some tweets mentioned more than one type of claim of fraud. We had a similar observation for other categories like Equipment, and Jurisdiction.

The authorship data set consists of tweets with multiclass labels associated with the tweet author, $human, llama, falcon, mistral$. Table 4 shows the number of instances for each of the labels in the training and the test sets.

Dataset	llama	falcon	mistral	human	Total
Train	2000	2000	2000	2000	8000
Test	365	310	350	875	1900

Table 4: Label distribution for instances in the training and test datasets.

6 Claim Characterization

In this section, we report the performance of multiple LLMs in extracting the taxonomy attributes from the ElectAI dataset.

6.1 Large Language Models

Recently, LLMs have been widely used for several Natural Language Processing (NLP) tasks, including text generation and text classification. We use the three LLMs described in Section 5.1 as well as the Flan-T5 model described below, for claim understanding (extraction of taxonomy attributes).

Flan-T5-XL (Chung et al., 2022) henceforth Flan-T5, is a language model based on the T5 (Raffel et al., 2020) model, which is a Text-to-Text transformer model. This model was fine-tuned for better zero-shot and few-shot learning for over 1000 different tasks. The model is one of the few LLMs with support for languages other than English.

6.2 Approach

We approach the election claim understanding task as a question-answering (Q&A) task, using LLMs in an in-context learning setting. In this Q&A setting, the input prompt is a question focused on a single taxonomy attribute, along with the target tweet, and the output is a yes/no response, indicating whether the specific attribute is relevant to the tweet. The set of questions that span the taxonomy attributes is shown in Table 5. In initial experimentation, we found that the taxonomy attributes associated with Other, like Jurisdiction-Other, were ambiguous and challenging for the models to extract. In mapping the taxonomy attributes to the questions in Table 5, we omitted the attributes associated with Other and instead included the associated parent label. For example, Jurisdiction-Other was replaced with a broader Jurisdiction label indicating whether any jurisdiction information is present (yes vs. no). We explored zero-shot learning, where no example inputs and outputs are provided in the prompt, and few-shot learning, where 3-5 input-output pairs are provided as examples in the prompt. However, the few-shot experimentation did not improve performance, so we only present the zero-shot results. Information extraction performance is evaluated using the F1-score. The following prompt was used to elicit model predictions.

Tweet: {tweet}
Answer the following question with a yes or a no.
Question: {question}
Answer:

ID	Question
1	Does the tweet mention a jurisdiction?
2	Does the tweet mention a state election?
3	Does the tweet mention a county election?
4	Does the tweet mention the federal election?
5	Does the tweet mention any election equipment?
6	Does the tweet mention electronic voting equipment machines?
7	Does the tweet mention ballots or related equipment?
8	Does the tweet mention any election-related process?
9	Does the tweet mention the vote counting process?
10	Does the tweet mention any election-related claim of fraud?
11	Does the tweet mention corruption in elections?
12	Does the tweet mention illegal voting?

Table 5: The input questions to the zero-shot prompts. Each question is based on an associated label from the taxonomy.

6.3 Results and Discussion

Table 6 shows the F1 scores for the LLMs evaluated using the annotated benchmark dataset.

Label	F1			
	Llama-2	**Mistral**	**falcon**	**Flan-T5**
Jurisdiction	0.828	0.851	0.554	0.898
Jurisdiction - State	0.865	0.923	0.602	0.933
Jurisdiction - County	0.562	0.852	0.705	0.824
Jurisdiction - Federal	0.239	0.290	0.263	0.277
Equipment	0.526	0.429	0.427	0.524
Equipment - Machines	0.720	0.904	0.692	0.870
Equipment - Ballots	0.510	0.609	0.540	0.549
Processes	0.598	0.606	0.398	0.607
Processes - Vote Counting	0.539	0.644	0.536	0.419
Claim of Fraud	0.897	0.758	0.513	0.917
Claim of Fraud - Corruption	0.452	0.523	0.497	0.544
Claim of Fraud - Illegal Voting	0.577	0.611	0.317	0.571
Overall	0.610	0.667	0.504	0.661

Table 6: Zero-shot in-context learning results. We report the F1 scores for individual labels as well as the macro-F1 score for each of the models

Mistral achieved the best overall performance at 0.667 F1, followed by the Flan-T5 model at 0.661 F1. falcon achieved the lowest performance among the evaluated models at 0.504 F1. We also show the F1 scores for each of the labels used. Apart from the falcon model, all the models had a better performance identifying the Jurisdiction and the Jurisdiction - State labels. Mistral and Flan-T5 models have a better performance for the Equipment - Machines label, as compared to the falcon and Llama-2 models. However, the performance of all the models in identifying the Jurisdiction-Federal label is low. Our analysis suggests that most tweets do not contain explicit

references to federal elections and that federal jurisdiction is typically implicit. Performance for labels like Equipment and Processes is low. Some tweets may mention equipment, like voting machines, ballots, or other election equipment. However, the models struggle to identify this information from a given tweet. The tweets are annotated for particular labels even if some information is implicit. Hashtags are also considered during the annotation process. We observe that some LLMs are unable to identify this implicit information. For example, if the word "ballots" is preceded by a hashtag ("#ballots"), the LLMs are unable to accurately identify the tweets referencing ballots. Another issue is the LLM's ability to understand and extract information from a given text. For example, the LLMs may label a tweet as mentioning a machine, even though voting machines are not mentioned in the tweet. Similarly, the LLMs may be unable to detect some specific terms and phrases like the name of a county if the word 'county' is not explicitly mentioned. These issues affect the performance of the LLMs, as evident from the results.

7 Authorship Attribution

In this section, we report on our evaluation of the performance of the Authorship Attribution Task, where the objective is to identify the author of the tweet, whether human or specific LLM.

7.1 Classification Models

We conducted experiments with various models as outlined below.

Random Forest (Breiman, 2001) is a supervised machine-learning algorithm based on decision trees. It aggregates the outputs of multiple decision trees to produce the final output. Our experimentation involves utilizing two types of input features: Term Frequency - Inverse Document Frequency (TF-IDF) and Word2Vec embeddings (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c), where we add the individual word vectors for each word in the sentence to output a sentence vector.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based model that has become ubiquitous in NLP. It has demonstrated state-of-the-art performance across various NLP tasks. In our study, we fine-tuned the BERT-based model with a multi-class classification objective. The BERT-base model is pre-trained on a substantial corpus of data, contributing to its robustness and effectiveness.

Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) is a variation on BERT that uses a more dynamic masking strategy, which improves feature learning. Similar to BERT, we fine-tune this model for multi-class classification.

7.2 Approach

We conducted multi-class classification to distinguish between human- and AI-generated tweets and resolve the specific LLM author. Three models, as described previously, were either trained or fine-tuned for this task. The Random Forest classifier was trained using two sets of features: TF-IDF and Word2Vec embeddings. We utilized grid search to determine the optimal values for the two hyperparameters, *N-estimators* and *Max Depth*. Specifically, we explored values of N-estimators=[50,100,200] and Max Depth=[20,40,50]. For the BERT and RoBERTa models, we employed the AdamW optimizer with a learning rate of 1e-5. The models underwent training for 3 epochs. Subsequently, all models were fine-tuned using the training data and evaluated using the test dataset. We report the F1 score for each model.

7.3 Turing Test

We performed a Turing test to evaluate how well AI-generated tweets mimicked human-generated tweets. During the second phase of annotations, the annotators were asked to label the tweets as human- or AI-generated. These annotations were compared with the actual labels for the tweets. The overall accuracy in identifying the human- and AI-generated tweets was 36.5%. We also calculated how accurately the annotators identified only the AI-generated tweets. We observed that only about 19% of the annotators were able to identify the AI-generated tweets. These results indicate that the AI-generated tweets are similar to human-written tweets. We observe specific patterns in the tweets generated by specific LLMs. Identifying these patterns and writing styles becomes straightforward if these tweets are presented consecutively. However, it may be difficult to identify these patterns if the tweets are presented randomly, especially with the human-written tweets included among the AI-generated tweets.

7.4 Results and Discussion

The best classifier for Random Forest with TF-IDF vectors had the N-estimators value of 200 with a Max Depth of 50, while the best classifier for Random Forest with Word2Vec Embeddings had N-estimators=200 and a Max Depth of 40. Both the BERT and the RoBERTa models achieved the maximum training F1 score at epoch 2.

Table 7 shows the F1 scores for each of the individual labels as well as the overall model. The transformer-based models, BERT and RoBERTa, achieve higher performance than the Random Forest models overall. All models have a superior performance identifying the human-generated tweets achieving a higher F1-score as compared to the tweets generated using LLMs. Moreover, these high performance of models, especially the transformer-based models can be attributed to the similarity between the tweets generated using several LLMs. We observe that tweets generated by each LLM tend to be homogeneous in sentence structure, vocabulary, phrasing, and other linguistic features. This makes it easier for the machine learning models to classify the tweets for the authorship attribution task, as they are exposed to a large number of tweets from each LLM in training.

To better understand the linguistic similarities and differences across the human and AI-generated tweets, we clustered the tweets based on the embeddings used to train the models, as shown in Figure 3. For the transformer models, we concatenated the last four hidden layers of the pre-trained BERT and RoBERTa models to create sentence embeddings, based on prior work (Devlin et al., 2019). Note that these embeddings were generated using the pre-trained models, not the fine-tuned versions from our experimentation. We use t-SNE (Van der Maaten and Hinton, 2008) to convert the high-dimensional embeddings to two-dimensional vectors. These are then plotted based on their respective author labels. We observe that clusters are formed based on the LLMs used to generate the tweets and these clusters are separable for the transformer-based models. Similarly, for the TF-IDF embeddings, the human-generated tweets form a separate cluster and are separable from the others. However, Word2Vec embeddings form small clusters, where the authors are not as easily separated. The separability of the clusters using the TF-IDF, BERT, and RoBERTa embeddings demonstrates that each LLM uses a distinct *voice*, which makes it relatively easy to resolve the tweet author. There are inherent patterns in the tweets, which are visually indistinguishable from humans, however, machine-learning models can identify these patterns and isolate the tweets. However, these patterns become increasingly apparent when humans are provided with a greater quantity of tweets generated by a specific LLM.

8 Conclusion and Future Work

We presented a novel taxonomy developed to further our understanding of election claims in social media. We used this taxonomy to curate ElectAI, the first benchmark dataset for election claim understanding containing tweets generated by humans and AI. Using ElectAI, we performed various experiments on claim characterization – answering **RQ1** – and authorship attribution of human- vs. AI-generated tweets – answering **RQ2**.

With respect to **RQ1**, our results showed that although LLMs have achieved state-of-the-art performance on several NLP tasks, these models have a moderate performance on claim understanding in an in-context learning setting. Among the models tested, we showed that Mistral performed best with an overall performance of 0.667 F1 score while falcon performed worst at 0.504 F1 score. With respect to **RQ2**, we showed that humans perform very poorly in discriminating between human- and AI-generated content. The results of a Turing Test show that humans can correctly discriminate human- from AI-generated with just over 36% accuracy. Computational models, on the other hand, perform very well on this task with a RoBERTa model achieving a 0.983 F1 score in identifying the authorship of tweets in a four-class experiment with tweets generated by humans and three different LLMs.

As part of our future work, we plan to annotate more instances of the ElectAI dataset using the proposed taxonomy and conduct experiments on a larger scale. We plan to evaluate various approaches that can potentially improve the performance of the models, such as few-shot learning, instruction fine-tuning, and chain-of-thought prompting. Finally, the work presented here is a first step toward identifying misinformation in election claims. We further plan to extend this work to identify misinformation and verify the claims using fact-verification approaches.

REFERENCES

Abilov, A., Hua, Y., Matatov, H., Amir, O., and Naaman, M. (2021). Voterfraud2020: a multi-modal dataset of election fraud claims on twitter.

Barbera, P. (2018). Explaining the spread of misinforma-

Model	Label				
	human	**llama**	**falcon**	**mistral**	**Overall**
Random Forest + TF-IDF	0.982	0.638	0.752	0.861	0.837
Random Forest + Word2Vec Embeddings	0.926	0.518	0.580	0.617	0.660
BERT	0.991	0.964	0.975	0.953	0.971
RoBERTa	0.998	0.974	0.987	0.975	0.983

Table 7: Model performance for the Authorship Attribution task. We report the F1 scores for each label as well as the macro-F1 scores for each of the models.

(a) TF-IDF Vectors

(b) Word2Vec Embeddings
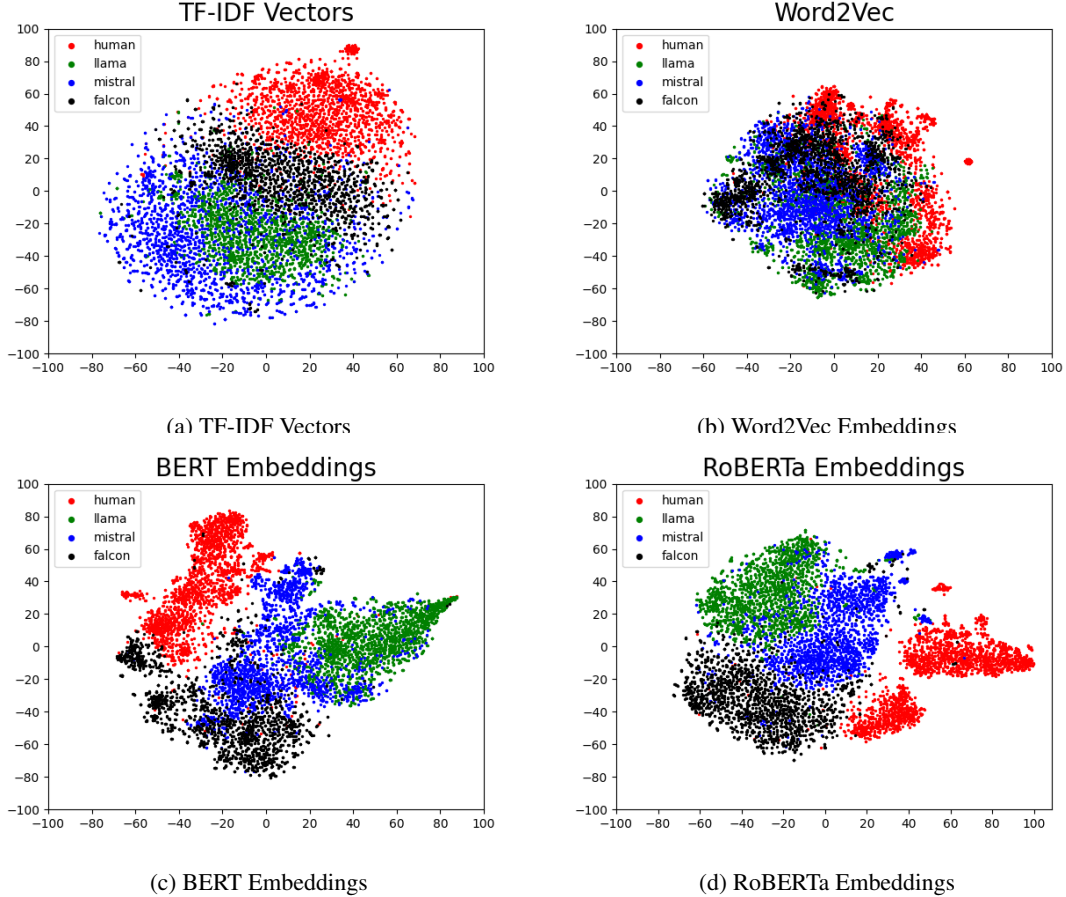
(c) BERT Embeddings

(d) RoBERTa Embeddings

Figure 3: The clusters of the embeddings used to train the machine learning models for authorship attribution task

tion on social media: evidence from the 2016 us presidential election. In *Symposium: Fake News and the PoliticsofMisinformation. APSA*.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, J., Zhang, R., Guo, J., Fan, Y., and Cheng, X. (2022). Gere: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2184–2189.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. In *arXiv preprint arXiv:2210.11416*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Heaven, W. D. (2022). Why meta's latest large language model survived only three days online. *MIT Technology Review. Last accessed December*, 15:2022.

Hu, X., Hong, Z., Guo, Z., Wen, L., and Yu, P. (2023). Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2319–2323.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pre-training approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sheikh Ali, Z., Mansour, W., Haouari, F., Hasanain, M., Elsayed, T., and Al-Ali, A. (2023). Tahaqqaq: a real-time system for assisting twitter users in arabic claim verification. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 3170–3174.

Sun, Y., He, J., Lei, S., Cui, L., and Lu, C.-T. (2023). Med-mmhl: A multi-modal dataset for detecting human- and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., and Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt. *Entrepreneurial Business and Economics Review*, 11(2):7–24.

Yang, R., Ma, J., Lin, H., and Gao, W. (2022). A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1761–1772.

Yao, B. M., Shah, A., Sun, L., Cho, J.-H., and Huang, L. (2023). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.