Unsupervised Detection of Inauthentic Accounts on Reddit

Using Coordinated Behavioral Similarity Signals

Tevin Parathattal

Advisor: Dr. Keith Burghardt

University of North Carolina at Charlotte

Abstract

We have observed an increase in inauthentic behavior on social media, but while Twitter has undergone extensive research, Reddit has yet to be studied to the same extent. Our experiment took Luceri et al.'s (2024) framework for coordinated inauthenticity detection on Twitter and adapted it to the Reddit platform using the 2021 dataset from Academic Torrents. Over the course of one semester, we improved upon six versions of our pipeline until we settled on seven pairwise behavioral similarity signals combined with large language models. We performed our experiments on two different data sets: general scanning of 50,000 randomly sampled accounts and targeted detection of 2,561 finance subreddits. Fast reply, which is a signal that we developed for Reddit, proved to be the most reliable standalone signal, whereas text signals saturated after exceeding 50 percent flag rate. Reviewing the results of the LLM verification process, we observed that although the model managed to detect structurally different forms of inauthenticity, including referral spamming, repeating similar comments, the use of off-site scam facilities, and direct references to market manipulation activities, the LLM resulted in many false positives in cases of usage of community slang language sounding like promotional messages. The problem emerged specifically in finance communities during the retail trading period of January 2021 with the GameStop/Dogecoin phenomenon. Namely, the speech community used by these communities was linguistically identical to coordinated language. The main contribution of our study is the development of a Reddit-specific detection process and an analysis of its performance characteristics, along with observations on the capabilities of LLMs in detecting coordination.

1. INTRODUCTION

There are many subreddit communities that focus on specific topics, where users have built trust among themselves due to shared interests. The trust network allows anonymous users to establish credibility with one another, which also creates an opportunity for malicious individuals to impose their own agenda on unsuspecting users. This could result in financial losses, market manipulation, and artificially inflated asset prices. Finance-oriented subreddits are especially vulnerable because there are clear financial incentives to influence users to make certain investments.

The majority of studies about inauthentic coordinated behavior have been conducted in Twitter because data is easy to access through its open API, and it became a key player during significant events such as the 2016 U.S. presidential election [6]. It was also easier to gather data on Twitter than on Reddit, making Twitter the more important platform to study.

This paper adopts the method used by Luceri et al. (2024) for detecting inauthentic coordinated behavior in Twitter posts and applies it to Reddit. Based on the 2021 data we received from Academic Torrents, we created a seven-signal pipeline that includes signals reflecting Reddit's nature, such as fast-reply coordination, thread Jaccard overlap, semantic similarity using sentence transformers, and binary fraud detection using a language model. Rather than providing a static pipeline alongside our findings, we have presented the pipeline's evolution across many iterations, the outputs of those iterations, and the lessons learned from running it in two different scenarios. Related works are covered in Section 2, Methodology and Pipeline Iteration in Section 3, Results from our two runs in Section 4, Discussion about successes and failures in Section 5, Limitations and Future Work in Sections 6 and 7, respectively, and Conclusions in Section 8.

2. RELATED WORK

The study of inauthentic accounts has advanced considerably over the last decade, and most of it focuses on Twitter. Specifically, research by Ferrara et al. shows that bots could significantly impact public discourse on Twitter [1]. In another paper, Varol et al. created a classifier that can identify bot users based on their metadata, interaction structure, and content [2]. Also, research by Bessi and Ferrara demonstrates the effects of bots on online discussions during the 2016 U.S. presidential elections, showing that the majority of bots discussed election-related topics [6]. However, all three studies rely on the Twitter platform and cannot be used elsewhere.

More recent work has shifted the focus to identifying coordinated activity across several accounts, rather than classifying bots individually. For instance, Cresci provides an overview of bot detection research for the last ten years and claims that the field has been moving from individual-account classifiers to coordinated detection techniques [7]. Moreover, Pacheco et al. provide a framework for detecting coordinated action on social media networks, which treats coordination as the primary focus [8]. Luceri et al. (2024) have proposed an unsupervised method to detect coordinated inauthenticity on Twitter using pairwise comparisons across multiple signals, such as text content, links, and retweets [3]. To adapt their method to Reddit, we substitute Twitter-specific signals with their Reddit counterparts, for example, Jaccard similarity between threads and fast-reply coordination.

In terms of methods, Reimers and Gurevych (2019) demonstrated that transformer embeddings could encode meaning more effectively than bag-of-words techniques [4]. In our study, we leverage all-mpnet-base-v2 to generate the seventh signal, i.e., semantic similarity. Similarly,

Ezzeddine et al. (2023) found that inauthentic users can be classified based on behavioral content analysis [5]. We will follow a similar process to determine whether an account is fraudulent.

Most literature on coordination detection focuses on Twitter; however, Reddit has not been ignored in related research. Specifically, Weld et al. (2022) have established a taxonomy of values across Reddit communities, thereby providing insights into why certain communities are easier to manipulate than others [9].

3. METHODOLOGY

3.1 Dataset

To conduct our analysis, we used 2021 Reddit data available on Academic Torrents in the compressed ZST format, which includes posts and comments from throughout the year. The final pipeline was run in two different ways, using different sampling schemes.

The first setup involved a large-scale scan to help us understand how each signal operates. It entailed scanning the entire dataset, counting the number of posts by each user, and excluding users with fewer than 3 posts because there isn't enough behavioral signal for these low-activity accounts. We then took a random sample of 50,000 users and downloaded all their posts, yielding roughly 7.3 million comments and 1.6 million posts.

The second setup was a targeted run against finance-related subreddits, where the motivations behind manipulation would be strongest. The allowlist included 20 finance-related subreddits, such as wallstreetbets, CryptoCurrency, dogecoin, Bitcoin, stocks, investing, personalfinance, options, ValueInvesting, dividends, ETFs, SecurityAnalysis, pennystocks, RobinHood, btc,

ethereum, Forex, StockMarket, financialindependence, and one case of CryptoCurrency. This allowlist was used to filter posts during the data loading step, leaving only posts from the selected subreddits. After applying the minimum three-post filter, the resulting sample comprised 2,561 users who made 92,177 comments and 5,496 posts. The smaller sample allowed the similarity-based signal to run without memory problems that plagued previous runs against the larger set of 50,000 users.

3.2 Similarity Signals

Seven pairwise similarity signals were calculated for all accounts. For each signal, each account was compared with every other account, yielding a similarity score between 0 and 1.

Signals A and B are comment and submission TF-IDF, respectively. Here, the texts of an account were aggregated in one vector, and cosine similarity was measured between the vectors. Accounts with the same vocabulary have high similarity scores.

Signal C is a subreddit TF-IDF, which captures the degree of overlap between the subreddits in which two accounts contribute. As observed during the broader scan, this signal proved informative in its variation. However, in the finance-sampled scan, this signal collapsed because, by construction, all accounts were posted only in the same limited set of finance subreddits, and thus similarity scores converged to 1.0, indicating that the vast majority of individuals were problematic.

Signal D is co-URL TF-IDF, which examines whether two accounts use the same external links. We require accounts to have shared at least 5 URLs before applying them to this signal, to avoid the increased false positive rates that were set to 3 URLs previously, per our advisor's suggestion.

We further discovered and rectified a bug in the code: if most accounts in the sample set have zero URLs, their TF-IDF representations will be zero vectors; hence, two zero vectors will be considered maximally similar. To address this problem, we zeroed out all rows and columns for any account with fewer than the minimum number of URLs. After fixing this problem, the flagging rate of Signal D was lowered from 97.7 percent to 1.0 percent in the finance run. Despite the fix, upon reviewing the top URL pairs, we found out that the most common URL shared by accounts in the dataset was i.redd.it, the image-hosting site for Reddit, and occurs in all posts that embed images.

Signal E is the Jaccard similarity between threads, which represents the ratio of common threads that have been posted by both accounts. A thread is defined by the totality of the conversation connected to a post/submission. A high score indicates that the two accounts tend to appear together in conversations. Calculating signal E in a brute-force manner for 50,000 accounts is computationally unfeasible since it involves the comparison of every single account pair. An inverted index method was adopted to calculate Jaccard similarity only for pairs of accounts that share at least one thread, by creating a map of each thread and the set of accounts that commented on it. In addition to the above-mentioned approach, a hard limit was defined: threads with more than 50 commenters were excluded from the calculation of both the intersection and the union of accounts.

Fast reply, referred to as Signal F, identifies accounts that respond to the same piece of content within 10 seconds of each other repeatedly. The frequency of fast replies between pairs was calculated based on the maximum frequency recorded during a run. A fast reply is the most reliable coordination signal in our pipeline because it is difficult to occur randomly.

Signal G is semantic similarity with all-mpnet-base-v2 sentence transformer. Unlike TF-IDF, which compares exact matches of the words in texts, signal G calculates how semantically similar two accounts' postings are, irrespective of their vocabulary choice, by converting each account's posting into an embedding vector of 768 dimensions and calculating their cosine similarity.

3.3 Detection Method

In each signal, we filtered out suspicious accounts by cosine similarity. We used dynamic thresholds calculated through quantiles for both similarity measures. Specifically, we filtered out pairs of accounts that have a similarity score in the top 0.5 percent among those posting in the same primary subreddit, or in the top 5 percent when accounts are posting in different subreddits. In other words, having similar posts within a single subreddit is not very surprising, while posting in different subreddits is much more suspicious. Each signal will provide its own list of suspicious accounts, thus allowing us to assess each signal separately. However, we also generated a fused measure, a weighted combination of the seven similarity signals, which was used later in the network analysis.

Initially, we considered detecting suspicious accounts using eigenvector centrality because of its similarity to prior research. However, we decided to use cosine similarity because it showed greater stability across multiple runs.

3.4 LLM Categorization

Once we had run all seven signals, we used Claude Haiku to classify the most-flagged accounts based on their post content. The method we used to do so is to sort the accounts by signal counts and by the number of posts they have made to break ties, taking the top 50 accounts after each run and submitting their content to the large language model, asking for the label of the account

being classified as fraudulent or non-fraudulent with confidence level and reasons of one-two sentences long. The definition of fraud was expanded to include coordinated price manipulation, pump-and-dump, paid shilling, crypto scams, bots, and organized inauthentic activity.

The LLM verification step serves as an additional, explainable layer on top of our unsupervised pipeline, allowing us to audit any flagged accounts. For this experiment, we settled on a binary fraud/not-fraud prompt for both reportable runs so that the outputs would be comparable. An earlier exploratory configuration included a 3-way organic/shill/pump-and-dump prompt and resulted in different fraud rates.

3.5 Pipeline Iteration

There were six iterations of the pipeline in total during the semester, each uncovering what was not seen in the previous one and arguing that this approach is more responsible than presenting a single static pipeline.

In Version 1, eigenvector centrality was used for detection, and there were five signals (comment TF-IDF, submission TF-IDF, subreddit TF-IDF, co-URL TF-IDF, thread Jaccard). We tried this out with synthetic data initially and later on with real data from 2021 after obtaining the IRB permission.

Version 2 did away with eigenvector centrality, replacing it with cosine similarity, added a fast-reply signal, and implemented dynamic, quantile-based thresholds. At this stage, we can say the pipeline is ready for analysis.

The third version established the subreddit-dependent threshold procedure and an output that captures the shared threads for each flagged pair, using thread Jaccard.

In the fourth version, semantic similarity was introduced using the all-mpnet-base-v2 sentence transformer, first-category classification from LLMs was added, and the finance filter was enabled.

The fifth iteration of the model removed the finance-related filter to examine the entire population unfiltered, changed the LLM prompt to be binary, and implemented the optimized index for the thread Jaccard calculation, with a maximum of 50 commenters. The calculation of semantic similarity exceeded available RAM for the 50,000-account sample, and the binary LLM ran inaccurately.

The sixth version corrected the problem with the empty URL vector in the co-URL function, computed semantic similarities across the full 50,000 samples, and produced the first proper binary LLM run. This was the full scan.

The seventh version applied the finance-related filter and used the sixth version's code without further changes. Both scans were produced using the same codebase to eliminate implementation drift when comparing the results.

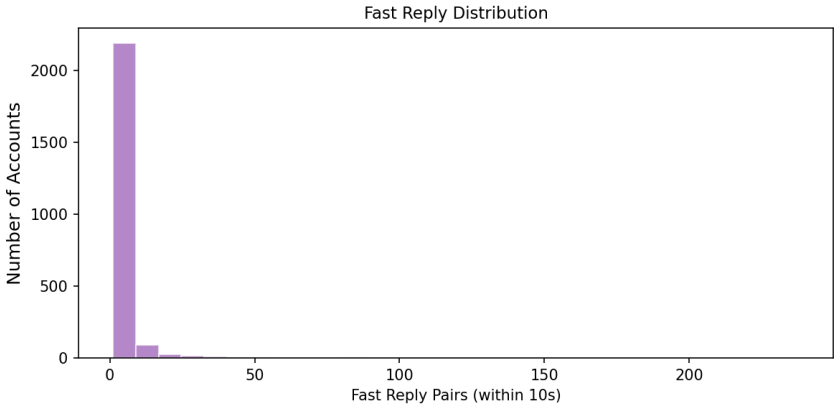
4. RESULTS

We present results from the broad scan (4.1), the finance-focused run (4.2), and a closer examination of what the LLM verification step caught and what it missed (4.3).

4.1 Broad Scan: 50,000 Accounts

signal	quantile	threshold	n_flagged	pct_flagged
A: Comment TF-IDF	0.99	0.0973	46236	95.7
A: Comment TF-IDF	0.95	0.0744	46236	95.7
A: Comment TF-IDF	0.9	0.064	46236	95.7
B: Submission TF-IDF	0.99	0.1068	35847	74.2
B: Submission TF-IDF	0.95	0.0683	35847	74.2
B: Submission TF-IDF	0.9	0.0548	35847	74.2
C: Subreddit TF-IDF	0.99	0.4338	41028	84.9
C: Subreddit TF-IDF	0.95	0.1358	46730	96.7
C: Subreddit TF-IDF	0.9	0.0693	47213	97.7
D: Co-URL TF-IDF	0.99	1	3953	8.2
D: Co-URL TF-IDF	0.95	1	3953	8.2
D: Co-URL TF-IDF	0.9	1	3953	8.2
E: Thread Jaccard	0.99	0.2	7396	15.3
E: Thread Jaccard	0.95	0.0588	18418	38.1
E: Thread Jaccard	0.9	0.0323	22736	47.1
F: Fast Reply	0.99	0.1111	40	0.1
F: Fast Reply	0.95	0.0417	173	0.4
F: Fast Reply	0.9	0.0278	341	0.7
G: Semantic	0.99	0.3819	48082	99.5
G: Semantic	0.95	0.2874	48304	100
G: Semantic	0.9	0.2421	48314	100

Table 1



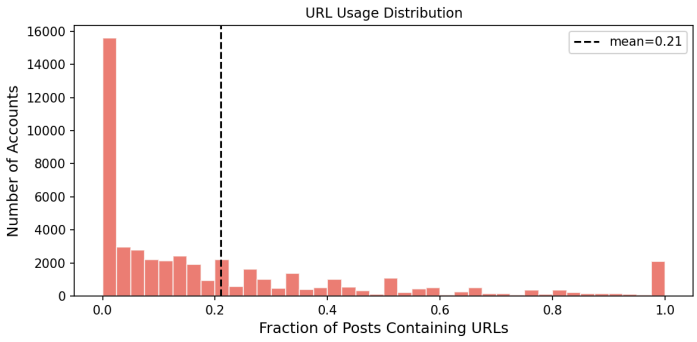
Fast Reply Distribution Figure

Table 1 shows the flagging rates of individual signals at the top 1%, 5%, and 10% thresholds for the broad scan.

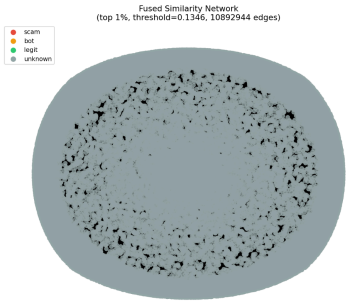
The fast reply (F) signal is the purest individual signal. It only flags 0.1% of accounts (40 out of 50,000) at the 1% threshold. The distribution of this signal has a sharp peak around zero and a thin tail, the profile of a discriminatory signal. The majority of accounts have between zero and five fast-reply pairs. The most extreme account in the sample has 238 fast-reply pairs with 41 different partners.

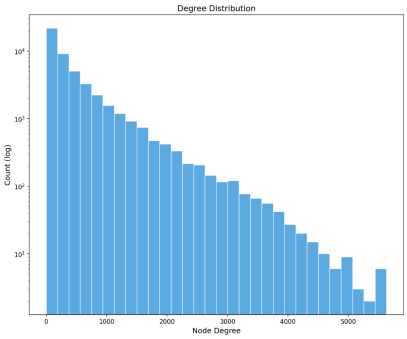
The thread Jaccard (E) signal flags 15.3% of accounts at the top 1% threshold. Many of these flags result from repeated co-presence in obscure threads after the megathread cap.

The text-based signals saturated. Comment TF-IDF (A) flags 95.7% of accounts at the top 1% threshold. The semantic similarity (G) signal flags 99.5%. The subreddit TF-IDF (C) signal flags 84.9% of accounts. These signals are not worthless, but they are not useful in isolation at the selected quantile thresholds. Rather, they serve as sieves that need to be crossed with other signals to generate candidates. Co-URL (D) caught 8.2 percent at the top 1 percent level post-fix. However, after reviewing the list of URLs, the issue persisted because i.redd.it was the prevailing URL pair.



The fused similarity network at the top 1 percent level included about 11 million edges spanning 50,000 nodes, with a degree





distribution that followed an approximate power-law pattern (most nodes had low degree, while a few had very large degrees exceeding 5,000 edges). Without any ground-truth data on the inauthentic accounts in the 2021 cohort, it is not possible to color-code the network by class; however, the degree distribution suggests the presence of many low-degree nodes alongside a handful of highly connected hub nodes within an environment of regular users.

In the LLM step on this particular run, using the top 50 most-flagged users, one fraud classification (BBMinus5, a confident Discord spam link in NSFW subreddits) and 49 non-fraud classifications were generated. These 49 non-fraud accounts were mostly high-volume accounts operating in the sports, anime, and AskReddit spaces. They had such a high volume that their accounts made it to the top flagged accounts, yet when the model checked for signs of suspicious activity, it classified them as organic users. This corroborates the signal saturation hypothesis stated earlier. At large scales, multi-signal flagging produces a list of candidates composed mostly of high-volume accounts. The LLM step was not unsuccessful, yet there wasn't much to be found.

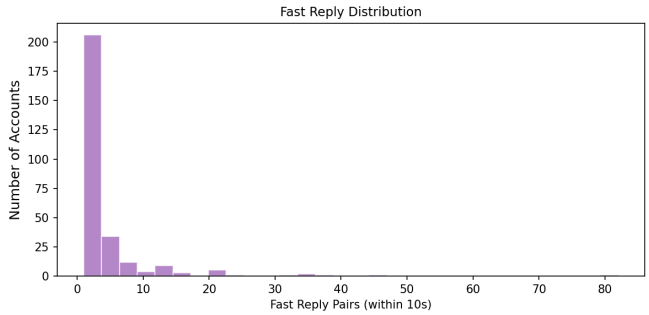
4.2 Finance-Focused Run: 2,561 Accounts

signal	quantile	threshold	n_flagged	pct_flagged
A: Comment TF-IDF	0.99	0.1016	2473	96.6
A: Comment TF-IDF	0.95	0.0755	2481	96.9
A: Comment TF-IDF	0.9	0.0634	2481	96.9
B: Submission TF-IDF	0.99	0.1576	1028	40.1
B: Submission TF-IDF	0.95	0.0947	1182	46.2
B: Submission TF-IDF	0.9	0.0745	1184	46.2
C: Subreddit TF-IDF	0.99	1	1579	61.7
C: Subreddit TF-IDF	0.95	1	1579	61.7
C: Subreddit TF-IDF	0.9	1	1579	61.7
D: Co-URL TF-IDF	0.99	1	26	1
D: Co-URL TF-IDF	0.95	1	26	1
D: Co-URL TF-IDF	0.9	1	26	1
E: Thread Jaccard	0.99	0.4	343	13.4
E: Thread Jaccard	0.95	0.2	1048	40.9
E: Thread Jaccard	0.9	0.1429	1315	51.3
F: Fast Reply	0.99	0.4967	9	0.4
F: Fast Reply	0.95	0.3333	26	1
F: Fast Reply	0.9	0.2222	65	2.5
G: Semantic	0.99	0.6329	1354	52.9
G: Semantic	0.95	0.5371	2044	79.8
G: Semantic	0.9	0.4845	2292	89.5

Table 2

The per-signal flagging rates at the top-1 %, 5%, and 10% thresholds for the finance-targeted experiment run are shown in Table 2.

Fast reply (F) flagged 0.4 percent of accounts (9 out of 2,561) at the top-1 percent threshold, consistent with its generalist approach. Thread Jaccard (E) flagged 13.4 percent. Both of these signals remained the cleanest stand-alone signals.



Fast Reply Distribution

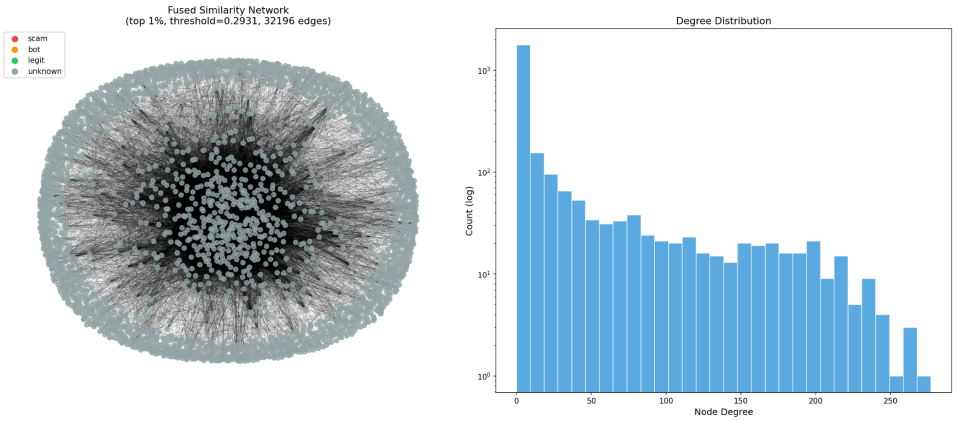
Three signals become saturated after applying the finance filter. Subreddit TF-IDF (C) reached the top-1 % threshold of 1.000; hence, there was no discrimination whatsoever. Similarly, co-URL (D) reached the top-1 % threshold of 1.000. Comment TF-IDF (A) flagged 96.6 percent of the accounts. In each case, the reason for this saturation was the same: filtering to a specific set of subreddits made it impossible for these signals to operate properly. All accounts in this experiment post within a very small set of subreddits, share the same most-frequent URL domain (i.redd.it) and finance-specific websites, and use finance-related terminology.

Signal B, using TF-IDF submission, and signal G, using semantic similarity, detected 40.1% and 52.9% of accounts, respectively, at the top 1 percentile threshold. Both results are quite high, but in this particular run, signal G exhibited a bell-shaped distribution centered at 0.28, which is characteristic of semantic embeddings. The signal functioned as it should, but financial accounts tend to cluster naturally in semantic space since they all discuss finance.

The pairwise comparison of signals demonstrated the following interesting result: signals F and E have no accounts in common in this particular run. The two most pure standalone signals capture two completely different groups of accounts in the finance configuration. This contradicts our expectations based on the literature on multiple signals, which holds that signals usually converge on some accounts. The reason could be that signal F detects coordinated

responses in real time (characteristic for bots/live events), while signal E detects coordinated presence in non-central threads (characteristic for niche communities).

The fused graph at the top 1 percent level has 32,196 edges among 2,561 nodes. This distribution



is more uniform than the previous one and shows no clear power-law curve. This suggests the graph represents a smaller, highly connected community, with most nodes having relatively similar connectivity levels, making it difficult to achieve high accuracy.

The result from our LLM algorithm for this trial, in which we selected the top 50 flagged accounts as inputs, identified 25 of the 50 (50 percent) as fraudulent. Out of these 25, 19 of them were found to be primarily active within r/dogecoin, while 6 of them were active within r/wallstreetbets. The remaining 25, which we labeled not fraudulent, showed activities within finance subreddits, again primarily in r/dogecoin and r/wallstreetbets.

4.3 What the LLM Caught and What It Missed

Examining the posts from the 25 users the LLM tagged as fraudulent, it can be seen that although the model accurately detected instances of inauthenticity, there were many false positives, where the basis for detection was that the account used the community's language.

Successful detections. The model was accurate at identifying structurally distinct inauthentic behaviors. We point out five accounts where posts show clear inauthenticity.

warg789 (r/dogecoin, 46 posts, all 5 discriminatory characteristics flagged). Made an explicit post containing a Binance referral link and a guide with seven steps telling the reader to "Create account on binance https://www.binance.com/en/register?ref=DDEG1XXH... buy doge... hold until the price go to mars... sell your doge coins for btc." An explicit referral link and a step-by-step guide constitute a structurally distinct inauthentic post.

shakyboy2828 (r/dogecoin, 37 posts, all 5 characteristics flagged). Five visible posts all have the same comment "Thank you. I appreciate your good advice." The LLM flagged 7 of this user's comments as verbatim copies.

ThrowRA-pi (r/wallstreetbets, 16 posts, all 5 signals flagged). Explicitly promotes wash trading activities: "buy at 310, sell at 320, buy at 330, sell at 340, etc. just passing the hot potato around." Also mentions the explicit planning of buy/sell activities to manipulate a stock's price: "We can all agree to do something. Let's all agree to place an order to purchase X stock at 30, then sell at 40. Rinse & repeat."

KaywanAhmed (r/dogecoin, 9 posts, all 5 signals flagged). Five out of five visible posts contain almost identical templates to advertise the HitBtc exchange ("Buy as much as possible on HitBtc as it is half price," "On HitBTc. Try to Buy doge as much as possible on HitBtc as it is half price there," and three more variants). Posting almost identical messages to promote a single exchange is highly indicative of paid affiliation or shilling.

Zayanza (r/dogecoin, 10 posts, all 5 signals flagged). Consistently requests payment methods and asks other members of the subreddits to acquire some amount of coin on behalf of the user. Asks for contact information, such as PayPal or Direct Messages.

What unites these five cases is that their fake behavior is fundamentally different from genuine community posting. It includes referral links with codes, repeated comments, open advocacy of

manipulation techniques, identical posts advertising a particular trade, and requests for payment outside the platform. These behaviors do not resemble those of real community users. The LLM successfully identified these fakes.

False positives. The LLM also flagged a number of accounts as fake when the sole indication was their use of community slang. Here are some typical examples of this false positive category.

kevingordillo94 (r/dogecoin, 44 posts): Examples: "Wooof you can hear from the moon 🚀🚀," "🚀Thousands of dollars in doge 🚀Let's GO 🔥🔥🔥🔥🚀🚀," "🚀🔥Nothing has more hype than doge coin." The LLM identified "aggressive hype language, unrealistic price predictions, repeated rocket/fire emojis." These types of posts were the standard on r/dogecoin during the January 2021 price run-up. There is nothing to suggest this account belongs to anyone other than an enthusiastic retail investor posting in the community's vernacular.

MrsLeviAckerman (r/dogecoin, 15 posts): Examples: "let's gooooo diamond paws," "I will hold for yearssss," "yah but It just dropped again bc ppl have paper paws lol." The LLM noted "repeated encouragement to 'hold' and buy, casual language mimicking coordinated retail investor messaging ('diamond paws,' 'leggooo')." These terms ("diamond hands," "paper hands," etc.) were widely used in r/wallstreetbets and r/dogecoin in early 2021. These phrases emerged organically within the community and are not associated with any coordination effort.

AyJayQz (r/wallstreetbets, 11 posts). Post examples are "Like and then reply so we can give eachother karma 🥂🙏😊" and "I'm sat with 18,000 DOGE - HOLD together we are all okay! 💎🙌." The LLM identified "explicit solicitations for karma, repetitive pump and dump terminology, and manipulative techniques." The first post is a joke. The latter is a dialect.

We found 5 solid catches, 3 probable catches based on citations we couldn’t verify from visible posts, and 17 possible catches where dialect was the sole basis of identification. This means that

actual precision among the top 50 in the finance run is somewhere between 10 percent (5 of 50 solid catches) and 16 percent (8 of 50 including both solid and probable catches), considerably less than the 50 percent indicated by the LLM's raw output.

The reason why this occurred. In January 2021, the organic users of r/wallstreetbets and r/dogecoin used pump-and-dump language because the GameStop short squeeze and the simultaneous rise of Dogecoin were driven by community action. These are real retail investors rather than mercenaries who use phrases like "to the moon," "diamond hands," "HOLD," and "rockets" as part of the lexicon of their communities during this period in history. The LLM, tasked to detect fraud, flagged the use of language as one of its criteria. As the dialect was nearly ubiquitous in the communities at this time, a large number of users in the dataset were flagged as fraudulent.

There was no problem with the pipeline per se because it successfully flagged similar accounts based on their behavior. Most of the similar accounts are from members of a community who speak in a particular way. However, it is the verification process through the LLM that made flagging an overcall, as the linguistic space had collapsed due to the finance filter.

5. DISCUSSION

5.1 What the Pipeline Captures

Of the seven signals used, only two carried over from Twitter to Reddit relatively seamlessly. Fast reply, which was designed specifically for Reddit, performed best among the seven signals and was the only one not susceptible to manipulation, generating a very difficult-to-mimic behavioral pattern. Thread Jaccard, another signal designed for Reddit, ranked second by purity, ranging from 9 to 15 percent. These two signals contain almost all of the pipeline's discriminatory power.

The other five signals were derived from text- or graph-based methods initially applied to Twitter coordination activities. However, they did not transfer smoothly into the Reddit context. Comment TF-IDF, subreddit TF-IDF, and semantic similarity all saturated above 50 percent in at least one of the two experiments. It should be noted that this does not necessarily mean that the methods behind these signals are inherently flawed. In reality, it is simply more difficult to separate coordination activity on Reddit due to the site's nature.

Two runs were required due to the nature of this experiment. While one run highlighted certain aspects of signal behavior, the other highlighted others. For instance, the first run identified the signals that actually worked at scale and their behavior in heterogeneous populations. The second one revealed the signals that break when any domain restriction is imposed.

5.2 The Filter and LLM Verification Tradeoff

The most intriguing result emerging from this analysis, methodologically, is the interaction between domain filtering and LLM verification, which is not well articulated in the existing literature.

Under the finance filter, the candidate set is filtered within a domain where there is an economic incentive for coordination. The finance filter successfully identifies true shills (KaywanAhmed promoting HitBtc), true wash-trading supporters (ThrowRA-pi), true scams (Zayanza), and bots (shakyboy2828). None of the unfiltered scans identified these accounts since the top-50 in unfiltered Reddit are high-volume sports and anime users.

However, the same filter that enhanced candidate identification eliminated linguistic features that the LLM would exploit to discriminate between genuine and fake activities. An LLM can distinguish between legitimate and fake activities in a heterogeneous community based on linguistic features such as vocabulary, sentence structure, and topic focus. However, when the

population is filtered by domain, every user will discuss similar topics, utilize similar vocabulary, and speak the same vernacular. The LLM relies on dialectal characteristics (emojis, slogans, hype talk) that are shared by both the fake accounts the filter is intended to exclude and authentic community members who have not been filtered out.

With historical context added to the dataset, leading to greater dialect richness, as seen in 2021 finance subreddits around the GameStop and Dogecoin events, this distinction becomes increasingly blurred. Our conclusion here is that LLM-based verification, beyond domain-based behavior filters, achieves high precision but yields high false-positive rates. A reviewer looking at the results without knowledge of the underlying data (a 50 percent fraud rate) will see one thing; a reviewer analyzing the content will come to a different conclusion. The difference should be reported clearly.

5.3 What Did Not Work

Three things failed in our pipeline in its current implementation. The co-URL Signal D is polluted by Reddit's native image host, i.redd.it, which has come to dominate the shared URL category. Despite having fixed the empty-vector problem, the signal is still mostly capturing image posting via shared URLs rather than actual coordinated linking. We need either a blacklist that excludes native Reddit hosts and all media sites, or a TF-IDF-like weighting based on rare domain names, to make this signal usable.

The subreddit TF-IDF signal C is impossible to compute in any filtered-domain run since filtering collapses the subreddit space first. It is interesting in the full-scan case but not in the finance one. Its use is conditional on the choice of run, which is hard to formulate in our unified language.

The first detection mechanism we considered was eigenvector centrality, which yielded inconsistent results in our tests. After applying a cosine similarity filter with quantile-based criteria (as advised by our mentor), we obtained more consistent results, and our signal analysis became easier to understand. We do not think that this was a fruitless step. Both options were tested to reach the final result.

6. LIMITATIONS

Without having ground truth for fraudulent accounts in our 2021 Reddit dataset, we cannot calculate precision or recall relative to a reference set. The "fraud rate" numbers we report here come from LLM classifications, which we demonstrated earlier to be highly untrustworthy in filtered-domain settings.

This study uses data from a single year, namely 2021. This year is special, as it saw the GameStop short squeeze, the Dogecoin surge, and an overall explosion in retail trading activity. Given this, it can be assumed that the dialect overcall issue is especially acute this year compared to other years. We are unsure whether our findings carry over to other years.

The finance-specific experiment involves only 20 subreddits, and the LLM classification step applies to no more than 50 users at a time. The 50-user-per-run limit for the latter step was arbitrary.

Our verification process for the LLM uses a single prompt and a single LLM model (Claude Haiku). Different prompts and LLM models would yield different results, including varying levels of fraudulent posts and different rates of false-positive detection. Our systematic manual check of the LLM's classifications relies on human subjectivity, as demonstrated by the classification of some posts as weak versus plausible/strong detection.

7. FUTURE WORK

The following list of experiments provides concrete steps to enhance the pipeline and research.

First, perform the LLM analysis step using a prompt that is more conservative in weighting dialect features (use of emojis, hype language, caps-locked enthusiasm) but specifically asks questions about templating, text repetition, off-platform solicitation, referral code use, and other structural indicators. Compare the detected fraud percentage to the current 50 percent.

Second, scale up the LLM analysis to include more than 50 users. The top 200, or even the top 500, will allow us to measure the model's precision and determine whether the bias in dialect classification is general or whether the top-flagged users have higher precision.

Third, repeat the experiment on Reddit data for another year, as 2021 has a significant anomaly introduced due to GameStop.

Fourth, test on different domains. Political, medical, and tech subspeak exhibit different dialectal overlaps and have distinct motivations for inauthentic actors. Cross-domain testing would highlight generalizable results from those domain-specific to finance.

Fifth, create a validated inauthentic user list through banned accounts from Reddit, suspended accounts from Reddit, and reported scam networks from Reddit. This replaces the reliance on ground truth generated by an LLM with true ground truth.

Sixth, incorporate temporal features. Coordination frequently forms in response to specific events (earnings calls, new tokens, etc.) and disbands once the event passes. Window-based signals can detect event-driven coordination that would otherwise be smoothed out by yearly aggregation.

Seventh, discard or revamp the saturated signals. Comment TF-IDF, subreddit TF-IDF, and semantic features (as implemented) have not demonstrated value at the quantile cut-offs. Discontinue, revamp with subreddit-relative scoring, or limit to the intersection only.

In the long term, a multi-platform comparison with Luceri et al.'s Twitter dataset will help distinguish between platform effects and methodological differences.

8. CONCLUSION

We adopted the coordinated inauthentic behavior detection framework used by Luceri et al. (2024) to detect Twitter accounts, modified it by running six pipeline iterations over one semester, and created two distinct pipeline runs: a general 50,000-account analysis and a finance-filtered 2,561-account analysis. We developed two new signals for detecting Reddit-specific structures (fast replies and thread Jaccard index with megathread filtering), incorporated semantic similarity using sentence-transformer algorithms, built the LLM verification pipeline, and corrected two bugs in the pipeline code.

Among the most crucial findings to emerge from this research on methodology is the realization that LLM verification following behavioral pattern recognition is viable for certain types of inauthentic activity but not for others. Instructive structural differences among inauthentic activity types, such as templated promotional tweets, repetitive identical comments, requests for off-platform payment, or explicit market-manipulation advocacy, are reliably detected. When the signal of inauthentic activity lies more in the linguistic than in the behavioral pattern, as with accounts posting in a linguistic style aligned with their community dialects at the highest point of community intensity, these will be incorrectly flagged by the LLM as organic community engagement. By filtering by domain, researchers help their LLM verifiers more effectively select

candidate instances of inauthenticity; however, the language feature space is narrowed, thereby making the issue worse rather than better.

References.

[1] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96-104.

[2] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[3] Luceri, L., Giordano, S., & Ferrara, E. (2024). Unmasking the Web of Deceit: Uncovering Coordinated Inauthentic Behavior with Graph Neural Networks. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*.

[4] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*.

[5] Ezzeddine, F., Luceri, L., Ayoub, O., Sbeity, I., Nogara, G., Ferrara, E., & Giordano, S. (2023). Exposing influence campaigns in the age of LLMs: A behavioral-based AI approach. *EPJ Data Science*, 12, 46.

[6] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday*, 21(11).

[7] Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10), 72-83.

[8] Pacheco, D., Hui, P. M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[9] Weld, G., Glenski, M., & Althoff, T. (2022). Making online communities "better": A taxonomy of community values on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*.