University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master degree in Computer Science

Automated Detection of Fraudulent Rental Posts on Facebook: Combining LLM with Open Source Intelligence

Supervisor

Prof. Alessandro Galeazzi

Co. Supervisors

Prof. Mauro Conti

Dr. Ying Yuan

Master's Candidate

Md Rubayet Afsan

2106016

Academic Year 2025–2026

Abstract

The growth of rental scams on social media, especially on Facebook, is becoming increasingly pervasive, which is particularly true among students searching for low-cost accommodation. Padova is a very competitive market for rent because of the University of Padova and its large population of Erasmus, local, and international students. Scammers target Facebook groups by posting fraudulent rental advertisements and comments under rent posts with unrealistically low prices, unclear details, or requests for private communication. They are often targeting some users who often overlook warning signs due to urgent housing needs. Existing detection methods are primarily focused on platforms like Craigslist, which have failed to address the real-time and community-based dynamics of Facebook public groups, leaving a critical gap in proactive scam identification. This thesis proposes a dual method for identifying the fraudulent rental listings using prompt engineering-based Large Language Models (LLMs) and Open Source Intelligence (OSINT) techniques. Using the OSINT tool Apify, we collected 2,189 rental posts from the public Facebook group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA." The dataset contained links, user IDs, advertisement contents, prices, images, addresses, and interactions (likes, comments, shares). The dataset was manually checked and labeled into categories: "Legitimate," "Suspicious," "Uncertain," and "Fake." Such classification was based on features such as price, images, address, user activity, interactions, and communications. We identified 32 fake posts through manual verification and generated 75 synthetic fake posts using a template-based method. This method analyzes suspicious and fake posts from the dataset and creates synthetic posts that are nearly identical to the original fake posts by maintaining the structure, incorporating vocabulary patterns, and adding statistically learned suspicious patterns, such as the lack of price, images, or detailed information. Two detection methods were developed and evaluated. One approach is a prompt engineering-based Large Language Model approach with the higher True Positive Rate of 0.969 on the original manually checked datasets and 0.935 after augmentation and the lower False Positive Rate of 0.254 on the original manually checked datasets and 0.254 after augmentation. The other approach is a Random Forest (RF) model with the lower True Positive Rate of 0.719 on the original manually checked datasets and 0.449 after augmentation and the higher False Positive Rate of 0.372 on the original manually checked datasets and 0.405 after augmentation. This shows that the prompt engineering-based Large Language Model outperformed the Random Forest model. Observations of unusual user behavior, including some suspicious activities, provided key insights into potential scam indicators. Examples include multiple comments offering rooms on several "Looking for rent" posts and the posting of many rental posts without prices or images. These examples were the primary reasons for suspicious activity. Scammers in particular posted advertisements and got banned and deleted from the group alongside their posts. They also targeted the comments section of the renters' posts to reach prospective victims through private emails or Messenger accounts, where they could bypass the platform's security measures. Engaging directly with suspected advertisers and commenters allowed us to expose prevalent deceptive practices, such as requesting payment of half a deposit using PayPal. This work highlights and tries to fill an important gap in real-time scam detection through manual post verification, which is enhanced by direct chatting with the suspicious advertisers and commenters.

Keywords

Rental Scams, Facebook Fraud Detection, Prompt Engineering based LLM, Synthetic data generation, Conversational Verification

Acknowledgments

I am thankful to my supervisor, Prof. Alessandro Galeazzi, and my co-supervisors, Prof. Mauro Conti and Dr. Ying Yuan. They gave me great advice and helpful feedback and kept encouraging me throughout this project. Their expertise and patience played an important role in this work.

I am also grateful to the University of Padova and the Department of Mathematics "Tullio Levi-Civita." They have provided an environment that gave me ideas and the tools I needed to do this research.

I would like to thank my friends and classmates in the Computer Science program. They have talked about different ideas and kept my spirits up. That made this journey fun and rewarding.

A special thanks goes to my family and friends. They never stopped believing in me and understood what I was going through during this academic journey. Their support gave me the strength to face the challenges and finish this thesis.

Finally, I want to thank the open-source community, the folks who created the Apify tool. Their work allowed me to collect the data that were at the heart of this study. This work would not have happened without the teamwork of all of the people I have mentioned.

Contents

Chapter 1

Introduction

The use of online platforms has increased the way people search for housing, especially students who are looking for budget-friendly rental options [Jia24]. This method also increased rental scams, with social media sites like Facebook becoming hotspots for such activities. Scammers take advantage of Facebook public rental groups by posting rental advertisements with some techniques, such as offering a bit lower price than usual, and some of them also write comments such as "available rooms" by asking to contact them in private under the posts section of "Looking for rooms." The use of this type of technique is increasing, and their ease of use in such public groups on the platform presents a daunting challenge for unaware users, as these pose both financially and emotionally damaging threats. There is an urgent need to devise new ways to detect and combat this type of fraud.

This thesis proposes a new hybrid framework that uses the functionalities of prompt engineering-based Large Language Model [Che25], Open Source Intelligence, and Facebook to detect fake rental listings. Concerning the Facebook public rental group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA," we have collected 2189 posts in total, with details such as group link, post links, user IDs, advertisement contents, prices, images, addresses, and interaction metrics (likes, comments, shares). The dataset was manually classified into categories as Legitimate, Suspicious, and Uncertain, based on an evaluation of features such as price, images, addresses, and user interactions. We have also labeled posts as 'Fake' after manually contacting scammers. Our methodology does not stop at the analysis stage but also includes interactive real-time conversational verification with suspected scammers as a way to prove fraudulent intention. This approach addresses the limitations of existing research on the Facebook platform, which relies on static datasets and lacks direct engagement with scammers.

To create a system that works well to detect scammers, we tried out and compared two models. These include a prompt engineering-based Large Language Model [Wan25b] and a Random Forest Model. The prompt engineering-based Large Language Model turned out to be stronger and outperformed the Random Forest model. The prompt engineering-based Large Language Model performed with the higher recall (True Positive Rate of 0.969 on the original manually checked datasets and 0.935 after augmentation) and lower (False Positive Rate of 0.254 on the original manually checked datasets and 0.254 after augmentation). The Random Forest model has lower recall (True Positive Rate of 0.719 on the original manually checked datasets and 0.449 after augmentation) and higher (False Positive Rate of 0.372 on the original manually checked datasets and 0.405 after augmentation). We have used a template-based method that gave us the augmentation of the dataset with 75 synthetic fake posts that further enriched our training data, allowing the model to generalize better across diverse scam scenarios. This work addresses current problems and challenges of the rental scams on Facebook. Our solution utilizes a prompt engineering-based Large Language Model to achieve a higher detection rate, significantly outperforming traditional methods such as the Random Forest model while providing real-time scam identification. This study aims to make the online rental markets safer and trustworthy.

The significance of this work lies in its contribution to filling a critical research gap [Mah23], while existing literature has explored rental scam detection on other platforms and Large Language Model-based approaches to misinformation detection. There is a lack of studies that have targeted the unique Facebook rental group ecosystem, where community interactions and real-time posting dynamics play a pivotal role. Our method is based on the integration of Open Source Intelligence for data collection and prompt engineering-based Large Language Model for analysis, which are combined with conversational verification and synthetic fake post data generation. It offers a comprehensive solution that adapts to the evolving tactics of scammers. Our findings reveal suspicious activities, such as multiple comments on "looking for rent" posts or the absence of essential listing details like prices and images. These are the key indicators of fraudulent activities. The ban and removal of scammer accounts from the group, along with their posts, underscore the reactive measures of the platform, which our proactive detection method aims to complement.

1.1 Background and Motivation

Online platforms have basically increased the housing rental market [Sin25]. It offers a convenient and accessible platform for people, especially students, to find accommodation. Facebook, in particular, has become one of the main places for rental listings. Large Facebook public groups like "AFFITTI PADOVA STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA" serve as local hubs where users can post, share, and negotiate rental opportunities in real time. This type of group promotes a perception of community. It allows members to access a wide range of listings to adapt to specific rental needs. However, this digital convenience has caused significant challenges, such as the proliferation of rental scams. That is, scammers exploit the trust and urgency of prospective tenants, turning these platforms into breeding grounds for fraudulent activities.

Rental scams on social networks have become a widely known form of cybercrime [Cha21]. They exploit advanced techniques rooted in social engineering and psychological manipulation strategies that use human emotions such as greed, fear, or the desire for quick solutions. Scammers craft misleading advertisements with tempting features such as unrealistically low prices or no prices, low-quality images or no images, and just a place name. These include not exact addresses, fake details about the property or landlord, and comments about availability on the renters' posts. These types of listings and comments are designed to attract renters to make quick decisions, which often pressures them to pay half of the deposit without verifying the authenticity of the property in person. The use of some techniques, such as offering a bit lower price than usual, has further expanded this threat. This enables scammers to create highly convincing content with minimal effort. For instance, a scammer might use some legitimate post structure to generate a detailed rental post for a made-up apartment in Padova. It is complete with professional-quality photos and a story about an urgent relocation with a fake address. Scammer techniques only disappear after collecting a half or full deposit via an unsecured payment method like PayPal. The accessibility and ease of use of these techniques pose an intimidating challenge for unsuspecting users. It results in significant financial losses, which often range from hundreds of euros based on half of the monthly rent or the full amount. It leads to intense emotional distress, including disappointment, stress, and loss of trust in online platforms.

The traditional and economical implications of rental scams are particularly noticeable in communities like Padova, where a large number of students depend on affordable housing to continue their education. Groups such as international students, first-time renters, Erasmus students, and others with limited experience in navigating online markets are especially unprotected. They are often lacking the resources or knowledge to avoid such fraud. These scams reduce the integrity of online rental markets, destroying users' confidence and creating challenges for platform administrators. Current platform responses, such as banning suspicious user accounts or removing fraudulent

posts, are largely insufficient. This is because scammers can easily create new profiles or shift to private communication channels like email or Messenger to avoid detection. This system dynamically highlights the urgent need for proactive and innovative detection methods. This can keep pace with the evolving strategy of scammers and protect vulnerable users in real time.

The academic and practical study of rental scams on social media has been limited by several key gaps in the research area [Wan25a]. Several studies on rental fraud exist, but they mostly focus on other platforms such as Craigslist or HousingAnywhere. This leaves Facebook's unique nature of rental groups unexplored. For example, Belloni's work on scam detection in online housing offers utilized a proprietary dataset of listings and addresses, which accomplished practical success through model ensembles. However, it struggled with dataset drift as user behaviors shifted over time. As with Van Der Zee et al.'s study of Craigslist rental scams, it provided valuable insights into scammer psychology, such as the use of sympathy, authority, or greed to manipulate victims. It was constrained by platform-specific policies and lacked real-time engagement with perpetrators. These studies are informative but do not address the specific dynamics of Facebook rental groups, where community interactions (e.g., comments, likes, shares) and real-time posting create a challenging environment for fraud detection. The absence of publicly available datasets specific to Facebook rental scams has obstructed the development of targeted detection frameworks. That is why researchers are choosing proprietary or platform-restricted data that may not generalize to other contexts.

Some limitations of existing research are focusing on other platforms and the lack of availability of Facebook datasets, which fails to detect many scams on Facebook and also on other social media. Traditional detection methods, such as rule-based systems, are effective in identifying obvious anomalies such as listings without addresses or prices, but struggle with the refined AI-generated content that scammers now employ. Mainly, the lack of direct interaction with the scammers in these approaches limits their ability to validate fraudulent activity. This gap is particularly significant given the rise of technologically assisted scams. This introduces complexities that traditional models are not the only method to handle. Theoretical studies, such as Reid's exploration of GenAI risks on sharing economy platforms, underscore the need for detection systems that can adapt to these emerging threats, but practical implementations remain scarce.

The motivation for this study arises from the crucial need to address these gaps and develop a robust, also real-time detection framework for rental scams on Facebook. By focusing on the large Facebook public group of Padova, this study leverages the unique dynamics of rental communities. Our use of Open Source Intelligence for data collection and prompt engineering based Large Language Model for analysis is combined with direct conversational verification. This offers a great approach that overcomes the limitations of fixed dataset analysis. This work is driven by a commitment to protect users, particularly students in Padova, who are facing significant risks from rental fraud. This study responds to an escalating threat with a timely and forward-looking solution, addressing both immediate challenges and long-term needs in the digital rental ecosystem.

1.2 Problem Statement

Nowadays, the rental scams on social media platforms like Facebook pose a significant challenge to users, especially students who are seeking affordable housing in Padova using Facebook public groups. Scammers use these platforms by posting unreasonable rental advertisements and several comments on renters' posts to target renters. Scammers are often targeting students who are not able to visit in person. They are also pretending to not be in the city and that there is no need to visit to rent. These scams lead to considerable financial losses through fraudulent payments and cause emotional distress by disrupting housing plans and undermining trust in online rental systems.

The problem is getting worse due to the limitations of platform-level responses. Those actions are taken by group admins, such as account suspensions or content removal, after reports are filed. These actions, while intended to mitigate damage, are often delayed due to manual review processes and fail to prevent scammers from entering the group again with new identities or shifting to off-platform communication channels. Whether it is encrypted messaging apps or email to finalize fraudulent transactions. This persistence poses significant challenges in regions like Padova, where high rental demand creates urgency that leads users, such as students, to overlook red flags. It is increasing their exposure to sophisticated scams. The absence of a systematic, real-time detection framework tailored to Facebook's unique ecosystem leaves users without adequate protection. It keeps going in a cycle of vulnerability and loss. The lack of integration between platform data and external validation methods hinders the ability to cross-reference suspicious activities. Such as repeated account creation or inconsistent user histories, which could serve as early warning signs.

This study tackles a key gap in the current scene by highlighting the need to develop a custom-built detection approach for the complex world of rental scams on Facebook. To address these gaps, this research develops a proactive detection framework that combines Open Source Intelligence with a prompt engineering-based Large Language Model. Our method specifically targets the evolving techniques of scammers on Facebook. Such as the misuse of community trust and requests for off-platform payments. The proposed LLM detector significantly mitigates the problem by achieving a higher recall rate, enabling real-time identification of fraudulent listings that evade conventional, slow-reacting platform tools.

1.3 Research Objectives

The objective of this research is to develop an advanced fraud detection system for Facebook rental scams through a systematic four-phase approach. First, we planned to collect raw JSONL data from targeted large Facebook public groups using Apify. Because there are no publicly available datasets for this specific study, we collected and curated our own. Then we needed to transform this into a structured Excel dataset with comprehensive metadata, including post content, prices, images, addresses, and user interactions. Second, we tried to implement a rigorous labeling protocol, where each post was manually classified as Legitimate, Suspicious, Uncertain, or Fake based on defined criteria. Then 32 fraudulent posts were confirmed through real-time conversations with advertisers. Third, to address class imbalance, we planned to augment synthetic fake posts using a template-based method. This system analyzes suspicious and fake posts from the dataset to extract full post templates. It is preserving their structure and semantics. The generator creates synthetic posts almost similar to real posts by maintaining the structure of the original post. It is incorporating vocabulary patterns and adding statistically learned suspicious patterns. Finally, we planned to develop and evaluate two detection models, training a prompt engineering-based Large Language Model specifically for scam identification. While comparing its performance against a supervised machine learning method, the Random Forest model. The ultimate objective was to establish a hybrid detection framework that combines the strengths of conversational verification, synthetic data augmentation, and AI-powered analysis to improve real-time scam identification in social media rental markets.

1.4 Our Contribution

This research makes several significant contributions to the field of online fraud detection. It is particularly emphasized on combating rental scams on social media platforms, specifically on Facebook. Our primary contribution is the creation and forthcoming public release of the first comprehensive annotated dataset of Facebook public group rental advertisements based on Padova.

It is a valuable resource that was notably absent from the existing literature, as confirmed through our extensive review of available sources. This dataset comprises 2,189 carefully curated posts from Padova-focused housing groups, including 32 verified fraudulent cases identified through our innovative real-time conversational verification process with advertisers, and is complemented by 75 synthetic fake post samples. These are generated by a prompt engineering-based Large Language Model to address class imbalance.

The technical contribution of our work focused on the development of a prompt engineering-based Large Language Model, and a Random Forest Model. The prompt engineering-based Large Language Model turned out to be stronger and outperformed Random Forest model. The prompt engineering-based Large Language Model performed with the higher recall (True Positive Rate of 0.969 on the original manually checked datasets and 0.935 after augmentation) and lower (False Positive Rate of 0.254 on the original manually checked datasets and 0.254 after augmentation). The Random Forest model has a lower recall (True Positive Rate of 0.719 on the original manually checked datasets and 0.449 after augmentation) and a higher (False Positive Rate of 0.372 on the original manually checked datasets and 0.405 after augmentation). We have used a template-based method to get synthetic fake posts. This method analyzes suspicious and fake posts from the dataset to extract full post templates. It is preserving their structure and semantics. The generator creates synthetic posts almost similar to real posts by maintaining the structure of the original post, incorporating vocabulary patterns, and adding statistically learned suspicious patterns. Our implementation of direct-advertiser engagement establishes a new paradigm for ground truth validation in scam detection research. Although our analysis reveals critical insights into scammer tactics specific to Facebook's ecosystem, particularly the prevalent use of comment section targeting and private channel diversion.

Except for academic circles, these contributions have immediate practical value for platform moderators and policy makers. It provides both detection tools and actionable intelligence to combat evolving fraudulent practices. The public availability of our dataset will enable future research replication and extension. Although our framework's design ensures adaptability to other platforms and emerging scam typologies, particularly those leveraging advanced scam techniques. Together, these advances address critical gaps in both research and practice, offering comprehensive solutions to the growing challenge of online rental fraud.

1.5 Scope and Limitations

This research focused on detecting fraudulent rental advertisements from a Facebook group. We could only analyze 2,189 posts collected from December 3, 2024, to March 15, 2025. This is primarily from the Italian language group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA," with some English-language content included. In this case, our prompt engineering-based Large Language Model framework and Google Translate for manual verification effectively processed this mixed language dataset. It has several important limitations that define the boundaries of our work. The study findings are most applicable to student/worker housing markets in Padova, where the data was sourced, and may not fully generalize to other regions or rental sectors. Although our prompt engineering-based Large Language Model approach with the higher recall (True Positive Rate of 0.969 on the original manually checked datasets and 0.935 after augmentation) and lower (False Positive Rate of 0.254 on the original manually checked datasets and 0.254 after augmentation) outperformed the Random Forest model. Its effectiveness relies on textual and behavioral patterns specific to Facebook's group dynamics. This means results may not be directly used on other platforms like WhatsApp or dedicated rental websites.

This collection of datasets provides robust exposure to scam practices during this period, but emerging threats, particularly AI-generated scams, may require future model updates. The way we have used our translation tools enabled a broad coverage of Italian and English posts but subtle linguistic nuances. As an example, local slang or abbreviations occasionally required manual review. The manual verification process, though rigorous, was resource intensive, confirming 32 fraudulent posts through direct engagement. This is a conservative estimate that may not capture all scam activity.

This work deliberately concentrates on text-based fraud detection, leaving image/video analysis and cross-platform scam networks for future research. Despite these limitations, our framework provides a strong foundation for detecting rental scams in multilingual Facebook public groups, which have adaptable methodologies that can be expanded to other contexts. Future studies could build on this work by incorporating automated verification, regional price benchmarks, and temporal analysis to track evolving scam patterns. After clearly defining these scope boundaries, we ensure a transparent interpretation of the results while highlighting the pathways for continued innovation in fraud detection.

Chapter 2

Literature Review

The digital era is widespread with online rental fraud, which capitalizes on human weaknesses. It employs advanced techniques focused on social engineering and psychological manipulation strategy. That is, it utilizes human emotions such as greed, fear, or the desire for quick solutions. This section looks at the work on rental scam detection, social engineering, and the advancing border of AI-assisted fraud. It is alongside our own work, which aims to detect Facebook fake house rental advertisements through a hybrid prompt engineering-based Large Language Model and Open Source Intelligence framework. We did not find any publicly available datasets on this, so our method includes scraping posts from the Facebook public rental group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA." It starts with 2,189, using the Open Source Intelligence tool Apify. This was followed by manual labeling of the posts into categories such as Legitimate, Uncertain, Suspicious, and Fake. It is interacting with users assumed to be fake to validate if they are scammers and then training a prompt engineering-based Large Language Model for detection.

2.1 Rental Scam Detection

Belloni [Bel18] studied at HousingAnywhere for six months, concentrating on automatic scam detection on online housing listing platforms. The author used a collection of offers from three years and thousands of marked addresses. Using this dataset, he trained several machine learning classifiers to detect listing fraud. This dataset is not publicly available due to its proprietary nature. However, the model did not continue to perform well on fresh data because of dataset drift because user behavior shifted over time. To counteract this, five distinctive models were trained in different time periods and features, which were practically successful in day-to-day operations. However, this study highlights the importance of adapting to behavioral shifts. It lacks real-time interaction with scammers, a gap our approach fills through conversational verification.

Van Der Zee and colleagues [VCA19] looked at how scammers try to trick people on Craigslist for long-term rentals in the UK. They gathered 2,112 letting advertisements over three weeks and talked to scammers in 44 email chats. This dataset is not publicly available due to privacy and platform restrictions. They found that scammers often use tricks to play on people's feelings, such as asking for sympathy, acting like they have authority, or tempting people with greed. These tricks match ideas from Cialdini and Stajano-Wilson. Even though scammers might seem skilled at first, they often use pre-written emails and don't know much about local culture, which gives them away. This study gives us useful information about how scammers act. This helps us label things by hand and figure out how to talk to catch similar tricks on Facebook.

Park and his team [PMS18] did a thorough study of Craigslist rental scams. They used computer programs to spot scam campaigns and chat with scammers by collecting data from Craigslist rental listings. This dataset is not publicly available due to the platform policies. They found many different tricks, such as credit report scams. They also learned that many scam campaigns asked for credit card payments, suggesting that stopping these payments could help fix the problem. Craigslist took down less than half of the fishy listings that were found, showing how hard it is to keep these platforms clean. Although their way of using computers to talk to scammers is similar to what we are doing, our study looks at Facebook. This has not been studied as much. We are also using LLM to analyze text, which should help us catch scams better.

2.2 Generative AI and Emerging Threats

Reid [Rei23] looked at how GenAI could help scammers on platforms such as Airbnb. He showed that GenAI can create believable fake profiles to hide the identities of scammers. This theoretical study did not use any dataset. This new threat makes rental scams harder to spot. Our work tackles this head-on by training a Large Language Model to find fraud in Facebook rental posts. This is new compared to earlier studies that did not consider GenAI.

This emerging threat is evidenced by recent developments in various kinds of social media platforms. Scammers have used GenAI to generate realistic property images and doctored content [Jia24]. That is, scammers use sophisticated fraudulent communications [Che25]. The use of synthetic data generation techniques has been shown to improve scam strategies [Cha+21]. Advanced Large Language Model-based detection methods are being developed to counter these threats [Wan+23]. These advances complicate traditional detection efforts in social media contexts [Mah23]. This illustrates the broadening scope of GenAI-assisted fraud. Our prompt engineering based Large Language Model approach aims to counter these threats by detecting fraudulent rental post content in Facebook groups.

2.3 Social Engineering in Scams

Chaganti et al. [Cha+21] studied recent trends in social engineering scams, including gift card scams. They suggested a threat model to map different types of scams. They noted tactics such as faking phone numbers and using events like COVID-19 to trick victims. Although they cover more ground, their insights on changing social engineering tricks are related to our study. We see similar tactics (such as creating urgency and fake trust) in Facebook rental scams [Bel18], but their work does not rely on rental fraud or Large Language Model-based detection methods [Lee+24]. Our approach shows that Large Language Models adaptively detect these scam techniques. It offers a proactive solution that addresses the dynamic nature of online scams. That is, traditional models may struggle to keep pace with [Wan+23]. This highlights the possibility of prompt engineering-based Large Language Model methods to bridge the gap left by earlier studies focused on broader scam categories.

2.4 LLM-Based Approaches to Misinformation Detection

Arowolo et al. [AMO22] proposed a gadget mastering approach to hit upon fake information on social media systems such as Facebook. It uses filtration and categorization algorithms to label content material as authentic or false. Shifting to the broader context of misinformation,

several research studies on fake news detection offer insights that are applicable to condo rip-off detection. Their observation highlights the assignment of deceptive content material. This mirrors the deceptive condo advertisements that we have addressed. However, their technique lacks the conversational verification and Large Language Model first-class tuning that are valuable to our method.

Wang et al. [Wan+23] introduced the FND-LLM Framework. This mixes small language fashions (SLMs) and Large Language Models for multimodal faux information detection. It integrates textual, visual, and cross-modal features to enhance its accuracy. Tested on datasets such as Weibo and Politifact, it achieved accuracy enhancements of up to 0.7%. While this framework leverages a Large Language Model for better detection, similar to our method, it makes a specialty of information articles as opposed to condominium commercials and does not include real-time person interaction. This is a key issue in our technique for validating rip-off conduct.

Li et al. [LZM23] advanced FactAgent, an agentic Large Language Model technique for faux news detection. It enables large language models to emulate human professional behavior in verifying information claims by breaking down the verification procedure into manageable substeps, without requiring specialized training. That is, using inner information and external equipment to evaluate veracity and give obvious reasons. Although FactAgent's adaptability and clarification capabilities are treasured, its awareness of information claims differs from our rental rip-off context. Nevertheless, its based workflow inspires our use of Large Language Model to systematically examine apartment put-up veracity.

Chen et al. [Che+23] proposed LEKD, a way to enhance the detection of fake information centered on text by combining SLM, outside understanding, and LLM through understanding distillation. It leverages a Large Language Model to generate supplementary expertise for training and uses graph-based total modules to clear up contradictions. It achieves stepped-forward overall performance over baselines. However, LEKD's use of LLMs for characteristic enhancement aligns with our technique. Its software and information content material and lack of real-time interaction differentiate it from our awareness of rental scams and conversational validation.

Papageorgiou et al. [PVC23] evaluated LLM-primarily based feature extraction for faux news detection. That is, introducing truth-based total datasets and exploring graph-based textual content representations. Their technique improves detection accuracy and interpretability in multiple datasets. This observer's emphasis on LLMs and interpretability resonates with our use of the Large Language Model for rental scam detection. Although our work uniquely applies these strategies to Facebook condominium posts and carries Open Source Intelligence pushed records series.

2.5 Prompt Engineering for Efficient Text Moderation

He et al. [He+24] conducted the first systematic evaluation of prompt learning with the Large Language Model to address toxic content online. Focuses on three main tasks: toxicity classification, toxic span detection, and detoxification. Their work demonstrates that prompt tuning optimization, only the prompt while keeping the LLM parameters frozen, can achieve comparable or even superior performance to traditional fine-tuned models. It is with significant reductions in training time and data requirements [PVC23]. For toxicity classification, prompt tuning improved F1 scores by over 10% compared to baselines. In toxic span detection, prompt tuning matched or outperformed state-of-the-art models such as SPAN-BERT but with much less computational cost [Lee+24]. For detoxification, prompt tuning effectively reduced toxicity scores. This was despite preserving semantic meaning and proving robust to adversarial perturbations [Wan+23]. This study highlights the adaptability and transferability of the prompt-tuned models. This can be generalized across datasets and requires fewer labeled samples. This is noticeable, making them

highly practical for dynamic online environments [LZM23]. This research underscores the value of prompt engineering as a lightweight, efficient, and robust approach to moderating toxic content on social networks and other online platforms. In alignment with these findings, our study also leveraged prompt engineering techniques to enhance detection accuracy and interpretability in the context of fake posts, suspicious posts, and scam identifications, which outperformed the Random Forest model [Che+23].

2.6 Multimodal Prompt Engineering for Evolving Threats

Lee et al. [Lee+24] explored the use of the multimodal Large Language Model for the detection of phishing webpages. This is done by leveraging prompt engineering to guide the Large Language Model. It is mainly focused on analyzing both visual (screenshots, logos, page themes) and textual (HTML) content aspects of web pages. Their two-phase system first identifies the brand a webpage is imitating and then cross-verifies it with the domain name to detect phishing attempts. This is unlike conventional computer vision-based models, which require continuous retraining on labeled datasets and maintenance of brand lists. The Large Language Model-based approach needs no labeled data and can flexibly interpret new phishing tactics. Evaluations on a newly collected dataset show that the Large Language Model system achieves high detection rates and precision. It outperforms state-of-the-art brand-based detectors and demonstrates resilience to adversarial attacks [He+24]. The system also provides interpretable explanations for its decisions. This is a feature that is lacking in previous solutions [Wan+23]. This work illustrates how prompt engineering can tackle the generalization and reasoning abilities of a multimodal Large Language Model. It addresses evolving threats such as phishing, without the overhead of traditional supervised learning pipelines [PVC23]. Similarly, our methodology incorporated prompt engineering strategies to guide the Large Language Model in identifying suspicious patterns and generating interpretable outputs, which is further validating the effectiveness of this approach in diverse online safety applications [LZM23].

2.7 Positioning and Novelty of the Research

Our method builds on these studies by focusing on Facebook rental scams. This is blending Open Source Intelligence for data collection (2,189 posts) with prompt engineering-based large language model-powered text analysis pipelines and real-time conversational verification. This is unlike Belloni, Park et al., and fake news studies, which rely on post hoc analysis. Our approach dynamically engages with scammers to confirm fraud, addressing gaps in real-time detection. Additionally, we tackle the emerging challenge of GenAI-assisted scams noted by Reid using fine-tuned prompt engineering based on a Large Language Model to detect fraud, a dimension not covered by prior works.

A key innovation of our work is the integration of prompt engineering inspired by recent advances [[He+24] [Lee+24]]. The purpose of this is to improve both detection accuracy and interpretability. We systematically extracted scam-indicative patterns from our dataset and designed adaptive prompts. That is, it guides the prompt engineering-based Large Language Model in identifying suspicious content. We used statistical weight learning that prioritized the most discriminative prompts. That is, allowing our system to adapt to new scam tactics and provide transparent reasoning for its decisions. This prompt engineering approach not only improves detection performance but also enables scalable, interpretable, and robust moderation in dynamic online environments. Our hybrid framework provides a comprehensive solution to the evolving landscape of rental scams and misinformation. It is combining real-time engagement and advanced prompt engineering based on a Large Language Model to address both current and emerging threats.

Chapter 3

Methodology

This study proposes a hybrid approach to detect fake house rental advertisements in Facebook groups. It is combined with Large Language Model and Open Source Intelligence techniques. The methodology is structured into six key phases. These are data collection, data processing, manual checking and data labeling, feature engineering, model development, and evaluation metrics. We will try to provide a comprehensive understanding of the process of each phase step by step.

3.1 Data Collection

Data collection is the foundational step of this study. We tried to find out if there were any existing datasets available based on Facebook rental groups, but we were not able to find any available datasets as required. Then we used the Open Source Intelligence tool Apify to collect raw datasets in a JSONL file, which are publicly available posts. We used Open Source Intelligence tools for Facebook data collection because it is the current best practice [Tea24; Res25b]. The problem we faced during collecting the dataset was not being able to scrape a large number of posts. In addition, sometimes some minor problems have been detected, such as missing some data. We manually checked each post to ensure that all the data is presented. Using this method, we were able to collect 2,189 rental posts from the Facebook public group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA." It consists of data from December 2024 to March 2025. This large group was chosen because it is a popular public group for students and workers in Padova who are seeking accommodations. This makes it a prime target for scammers.

The data collected includes a variety of features for each post. These include a group link for checking the group. The post links to check the post. The post text for the description of the rental offer. The user ID to check user activity. The price in euros and images, if any, as numbers. The address for specific location details, if provided, and interaction metrics such as likes, comments, and shares. These features were used to capture both the content of the posts and the social engagement they received, which could indicate authenticity or suspicion.

This JSONL sample data represents a single post collected using the Open Source Intelligence tool Apify. It shows a user who is looking for accommodation. It consists of fields such as user information, text content, and other details. There are no comments available from suspicious users. This aligns with our Uncertain post category.

```
{
  "facebookUrl": "https://www.facebook.com/groups/349362098841279/",
  "url": "https://www.facebook.com/groups/349362098841279/permalink
      /2112160992561372/",
  "time": "2025-03-15T13:36:13.000Z",
  "user": {
    "id": "pfbid0hrWm9RnSiCqFYz1uAQ3P58GTN3xVCrCq9YN9YUxLzKJEi4483m59eTeMafRCQ8abl",
    "name": "Francesco Piazza"
  },
  "text": "URGENTE\nCerco casa/appartamento/stanza per 2 persone.\nSiamo 2
      LAVORATORI con contratto a tempo INDETERMINATO",
  "topReactionsCount": 0,
  "feedbackId": "ZmVlZGJhY2s6MjExMjE2MDk5MjU2MTM3Mg==",
  "id": "UzpfSTEwMDAwOTA0MDI3NjkzMzpWSzoyMTEyMTYwOTkyNTYxMzcy",
  "legacyId": "2112160992561372",
  "likesCount": 0,
  "sharesCount": 0,
  "commentsCount": 0,
  "topComments": [],
  "facebookId": "349362098841279",
  "groupTitle": "AFFITTI PADOVA - studenti/lavoratori che vivono/vorrebbero vivere a
       Padova",
  "inputUrl": "https://www.facebook.com/groups/349362098841279/"
}
```


Figure 3.1: Example of a Uncertain "Looking for Rent" Post from the JSONL Datasets.

3.2 Data Processing

After successfully scraping 2,189 posts as raw datasets in a JSONL file from the Facebook public rental group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA." Using the Open Source Intelligence tool Apify, the dataset was subjected to a structured data processing pipeline to prepare it for analysis and modeling. The raw JSONL data was converted into an Excel file using Python code. It consists of standardized columns such as Advertisement Text, Price, Images, Address, and Interaction. These features have been chosen to evaluate possible scams. This approach is effective because it captures critical indicators of fraudulent behavior. The advertisement text reveals linguistic patterns and content quality that often signal scams, such as unclear descriptions. The price highlights unrealistic offers. The image assesses the presence or absence of authenticity of the place. The address verifies whether the location exists or not. The interaction reflects user engagement levels and shows low activity or suspicious comment patterns. This structured approach facilitated subsequent analysis and modeling.

This table presents a single rental post from the Excel dataset. It is showing the standardized columns (Advertisement Text, Price, Images, Address, Interaction). This is used for scam detection analysis. The advertisement text is detailed and specific, but in Italian. After translation into English, it says, "(Student accommodation available from the end of March (date negotiable). The cost is €300 per month, including condominium fees, excluding utilities (about €30-40 per month). The house is located in the city center, between Piazza dei Signori and the Arena Gardens. The accommodation is in a large double room shared with a very quiet and tidy woman. The apartment features a fully equipped kitchen, a large living room, and three bathrooms with a balcony off the living room and a terrace upstairs. The house is shared with two women and two men. There is room to store bikes in the building. Message me if you are interested! :)." The price is within a reasonable range, and five images are provided. The address is available but not exactly indicated, and the interaction is minimal.

Ad Text	Price	Images	Address	Interaction
Disponibile posto letto per studentessa da fine Marzo (data contrattabile)**. Il costo è di €300 al mese, spese di condominio incluse, bollette escluse (circa 30-40 euro al mese). La casa si trova in pieno centro, tra piazza dei signori e i giardini dell'arena. Il posto letto si trova in un'ampia camera doppia condivisa con una ragazza molto tranquilla e ordinata. L'appartamento presenta una cucina completa di tutto il necessario, un grande salotto e 3 bagni con un balcone sul lato del salotto e un terrazzo al piano superiore. La casa è condivisa con due ragazze e due ragazzi. Nel condominio c'è spazio per lasciare le bici. Scrivimi se sei interessata! :)	300€	5	in pieno centro	1 like

Table 3.1: Sample Rental Post Table.

For the Advertisement Text section, we have added full text from the posts. The Price section is based on the number and euro sign, for example, 300 €, or if there are no prices at all, then none. The Images section is based on how many images are posted in each post, for example 5, or if no images at all, then none. The Address section is based on the address mentioned in the post as an example, in via Luigi Pellizzo, in Center, in Zona Chiesanuova, etc., or if no address at all, then none. The interaction section is based on the number of likes and comments received on the post; for example, 4 likes, 2 comments, or if no likes/comments at all, then none.

3.3 Manual Checking and Data Labeling

Manual verification was one of the main parts of this study. It is ensuring high-quality ground truth labels for supervised learning and evaluation. Manual fraud review balances a perceptive element for high-accuracy detection[Res25a]. In the Excel file, we have added three more columns, which are Label, Fake, and Reasons. For the label section, we have tried to mainly label each post as Legitimate, Uncertain, or Suspicious. This is based on some criteria. The label "Legitimate" is based on detailed information, has enough or few images to know the property, has prices, has an exact address, and finally, less importantly, has good interactions or not. The label "Uncertain" is based mostly on "looking for rent posts." This does not have any comments from suspicious users offering rooms in the comment section suspiciously and asking for contact privately. The label "Suspicious" is based on, especially, no proper detailed information, no prices or suspiciously lower prices than normal, no images or just one, no proper address, or no address at all.

For the Fake section, we have tried to reach all suspicious advertisers and commenters by texting them. We have found some scammers who were banned by the group along with their posts for scam activities, and also some scammers were directly asking for half the deposit via PayPal from us without even offering to visit the place. So after confirming these scams, we have made a "Yes" label for the Fake section, and the rest is No.

For the Reasons section, we have given reasons for why we have labeled it and based it on certain criteria. As an example, "Legitimate": Reasonable price; Has images; Has address; Detailed information. "Uncertain": Looking for rent, no comments from suspicious users. "Suspicious": Several posts; Suspicious user activity; Contacted owner. "Fake (Yes)": The user who posted was banned and removed from the group along with their posts.

Ad Text	**Price**	**Images**	**Address**	**Interaction**	**Label**	**Fake**	**Reason**
2 posti liberi in camera doppia disponibili subito, ingresso immediato con la firma del contratto. Solo per studenti. Centro città. Affitto posto letto: 300€/mese Posizione: Corso Milano 122 Spese aggiuntive: Condominio: 89€/mese Bollette: 20-30€/mese	300€	5	Corso Milano 122	1 like	Suspicious	Yes	The user who posted was banned and removed from the group along with their posts;

Table 3.2: Sample Fake Rental Post Table.

This table presents a single "Fake" rental post from the Excel dataset. It shows the standardized columns (Advertisement Text, Price, Images, Address, Interaction, Label, Fake, and Reason). It is used for scam detection. The post is labeled Suspicious and confirmed Fake because of its multiple posts by the user and suspicious activities. The user was banned and removed from the group along with their posts.

Figure 3.2: Fake Post

In this figure, we can see an awareness post from a person who was scammed. He posted about how a scammer posts in the group and targets audiences. The advertisement text is in the Italian

language. After translation, the English advertisement text is ("For rent: Studio apartment in Padua - Ideal for students and young professionals. A studio apartment available for rent in Padua, perfect for students or workers with a contract. Located in a convenient area, close to all amenities: pharmacies. Rent: €600 WhatsApp:").

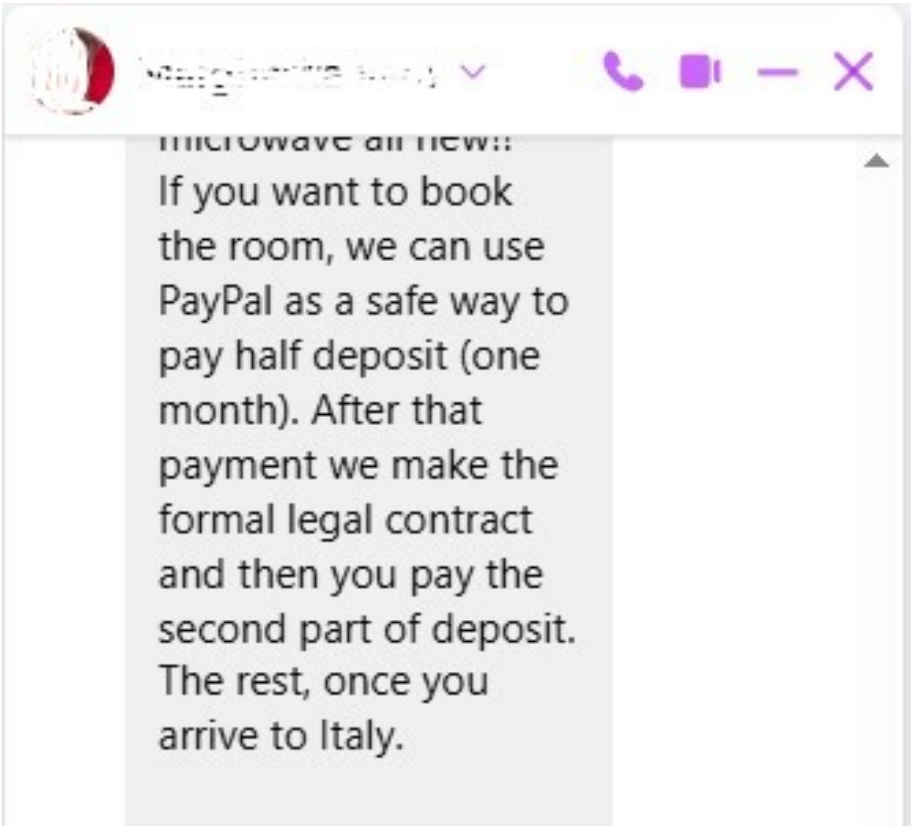

Figure 3.3: Scammer 1

In this figure, we can see that after contacting a suspicious user from a commenter in a "Looking for rent" post. The scammer is asking directly for half of the deposit by PayPal without visiting the place. In this type of case, after receiving half the deposit, the scammers do not keep in contact.

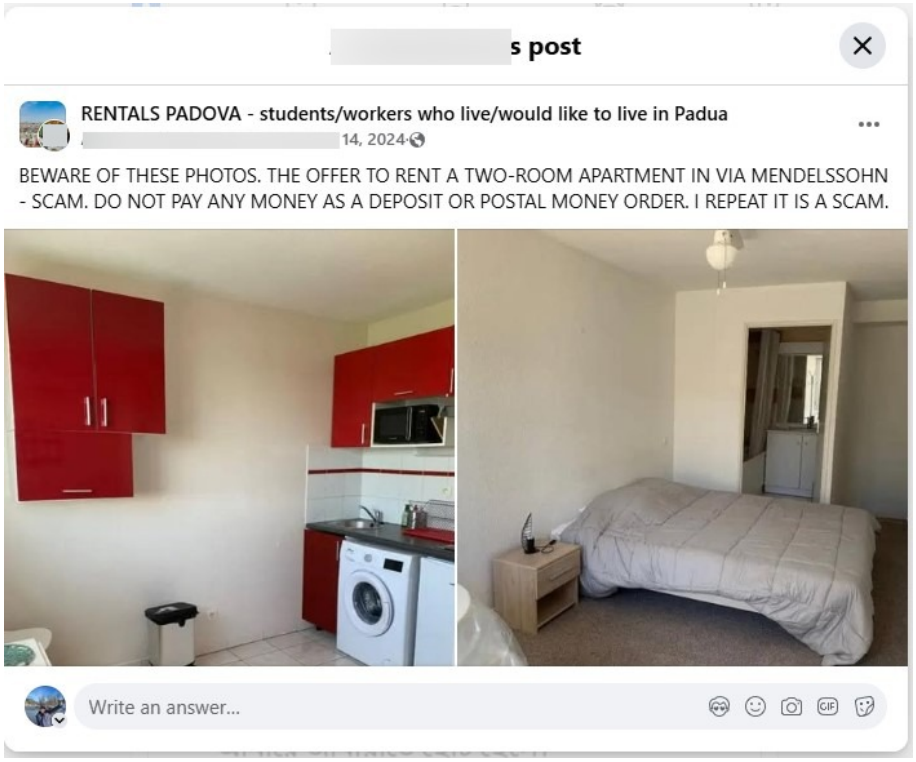

Figure 3.4: Beware of scams post 1

In this figure, we can see that a person was scammed and posted the exact post that was used to get victims. It is an awareness post. Here, it shows the kind of pictures and details that are used by scammers, and not to believe in such posts, where scammers are asking for pay money as a deposit or postal money order.

3.4 Feature Engineering

This study is designed to capture subtle patterns and signals that identify Legitimate, Uncertain, Suspicious, and Fake house rental advertisements in Facebook public groups. The approach is rooted in a hybrid framework that combines prompt engineering-based Large Language Model analysis, Open Source Intelligence-driven data collection, and manual conversational verification [Bah+16; Ike20].

Overview of Feature Creation: The dataset, as reflected in the Excel file, includes columns such as Advertisement Text, Price, Images, Address, Interaction, Label, Fake, and Reasons. Each of these fields provided the basis for feature derivation, with additional features engineered through text analysis and behavioral profiling.

Categories of Features:

- **Textual and Linguistic Features:**
  - Advertisement Length: Total number of words and characters in the post, reflecting simplicity.
  - Urgency: Binary indicator for phrases such as urgente, immediately, or similar.

  - Suspicious Keywords: Presence of scam-related terms such as "no visiting invitation," "send half deposit," "PayPal," and "caparra subito" (deposit immediately). This is identified through selected keyword lists and regular expressions.
  - Contact Information: Flags for private contacts such as phone numbers, email addresses, and WhatsApp within the advertisement text.
  - Address Mention: Whether the post provides an exact full address, a partial address, or no location details.
  - Grammatical Quality: Manual or automated spelling and grammar assessment, as scam posts often contain errors or nonstandard phrasing.

- **Numerical and Structural Features:**
  - Price: Advertised rental price, normalized to exact price in euros for consistency.
  - Image Count: Number of images attached to the post, as scams often lack images.
  - Interaction Metrics: Number of comments, likes, and shares to know community participation.

- **Behavioral and Contextual Features:**
  - User Reputation: Absolute variable based on the frequency of suspicious posts and any previous bans or warnings.
  - Posting Patterns: The time of day, day of week and the frequency of posts by the same user, capturing suspicious activity.
  - Response Behavior: For posts with conversational engagement, the features include willingness to provide more information and asking for a deposit before even visiting.

- **Manual and Derived Features:**
  - Label Reason Codes: Extracted from manual labeling notes, such as reasonable price, no images, suspicious user activity, etc.
  - Manual Flags: Indicators for posts flagged during manual review, such as a user banned after posting suspicious posts, a scammer confirmed via communication, a scam awareness post, and so on.
  - Synthetic Data Indicator: By focusing on originally collected suspicious posts. We have generated synthetically fake posts that are used for training and evaluation.

Feature Extraction Pipeline:

- Text Preprocessing: All advertisement texts are fully scrapped as they were posted. Even with emojis and special characters.
- Pattern Matching and Regular Expressions: Used to extract such as contact details, prices, suspicious keywords, and urgencies.
- Metadata Parsing: Image counts and interaction metrics are parsed from post-metadata or explicit dataset fields.
- Manual Annotation: Features such as grammatical quality and scammer confirmation are annotated during manual review and conversational verification.
- Missing Data Handling: Features with missing values, such as missing price or images, are imputed using None.

Feature Name	Type	Description	Example Value
Price	Numerical	Advertised price in euros	400
Image Count	Numerical	Number of images attached	3
Address Present	Boolean	Whether an address is provided	None
Urgent Language	Boolean	Use of urgent/pressure phrases	No
Suspicious Keyword	Boolean	Presence of scam-related terms	Yes
Contact Info	Boolean	Phone/email/link present in text	No
Interaction Count	Numerical	Number of comments/likes/shares	12
User Reputation	Absolute	Posting frequency, prior bans	warnings
Manual Flag	Boolean	Flagged by manual review or conversation	Yes
Synthetic Data	Boolean	Post is synthetic (augmentation)	No
Label Reason Code	Categorical	Reason for label assignment	Has images; Address

Table 3.3: Sample Feature Table.

This table outlines the feature set used in the scam detection model, detailing each feature's type, description, and example value.

This comprehensive and hybrid feature engineering process ensures that the detection framework is both interpretable and highly adapted to the evolving tactics of rental scammers on social platforms such as Facebook.

3.5 Model Development

This section presents the development of two complementary models for detecting fake house rental advertisements in Facebook public groups. The two approaches are a prompt engineering-based Large Language Model detector and a Random Forest model. Both approaches were designed to address the specific challenges of identifying fraudulent posts in the dataset collected and manually checked from the Facebook public group "AFFITTI PADOVA: STUDENTI LAVORATORI CHE VIVONO VORREBBERO VIVERE A PADOVA." Also, how using a prompt engineering-based Large Language Model detector can outperform the Random Forest model. This process involved careful feature preparation, model training, and repetitive improvement to optimize detection accuracy.

3.5.1 Prompt Engineering based LLM for Fake House Rental Advertisement Detector

The prompt engineering-based Large Language Model Fake Detector combines the deep semantic understanding of Large Language Model with a statistically driven weighting system to identify fraudulent rental posts. This approach relies on prompt engineering, where customized prompts guide the large language model in classifying posts based on their textual and contextual characteristics, thereby enabling the effective detection of fraudulent rental advertisements in Facebook groups. For further improvement of performance, a Statistical Weight Learning System was integrated. This approach dynamically assigns weights to suspicious patterns identified in the dataset, eliminating the need for manual weight assignment and enabling more adaptive and accurate fraud detection. Using real-time exception detection, this hybrid system not only reduces False Positives but also provides interpretable explanations for flagged posts, enhancing trust and usability in practical deployment scenarios.

Architecture and Components:

- **Statistical Weight Learner:** Executed as the StatisticalWeightLearner class. This module identifies and measures the importance of patterns such as urgent, private contact, immediately, WhatsApp, no images, etc. using correlation analysis, chi-square tests, information gain, and the Random Forest feature importance. The weights are combined using an ensemble approach of 30% correlation, 25% chi-square, 25% information gain, and 20% Random Forest importance to provide a balanced and robust evaluation of feature relevance. It captures both linear relationships, categorical dependencies, entropy reduction, and nonlinear interactions. By combining these, the system adapts to the dataset's variability, improving the ability of LLM to detect subtle fraud patterns in rental posts. The specific weights appear to have been empirically chosen to prioritize the foundational metrics (correlation). Equally balancing categorical entropy-based methods and slightly emphasizing model-specific importance. The tuning process eliminates manual intervention for better scalability and performance.
- **Prompt Engineering Framework:** The prompt engineering-based Large Language Model is prompted to analyze the full text of each post. It is incorporating the Advertisement text, Price, Images, Address, Interaction metrics, Labeling rationale, Fake, and Reasons. The prompts are designed to extract step-by-step reasoning, enabling the model to detect suspicious patterns and assign a probability score for fraudulence.
- **Iterative Weight Refinement:** Training is carried out in multiple iterations, such as 30, with a learning rate of 0.9. After each iteration, the weights are updated for misclassified posts of ± 0.1, allowing the model to adapt to subtle patterns and improve discrimination. The choice of 30 iterations was determined to ensure convergence in the 2,189 rental posts. It enables the model to refine its detection of subtle scam indicators without overfitting. This number balances training efficiency and accuracy. This aligns with the project's real-time detection objective. The learning rate of 0.9 has been used to enable aggressive weight updates. This allows for quick adaptation to the variation of the datasets, which includes synthetic data, and effectively captures dynamic features like Price and Suspicious keywords. This high rate moderated by ± 0.1 adjustments ensured stability. It has improved the performance of the Large Language Model over the Random Forest model. This addresses the need for strong scam identification in the competitive rental market in Padova.

Training Process: The model is trained based on a dataset of 2,189 real posts and augmented with 75 synthetic fake posts. This is generated by the template-based method from suspicious posts. Synthetic posts are designed to consider realistic scam patterns and address class imbalance. Each post is represented as a feature vector that indicates the presence and weight of suspicious patterns. A threshold of 0.15 is used to classify posts as fake (probability > 0.15) or Legitimate. The model reasoning steps are recorded for understandability, which provides transparency into classification decisions. The threshold of 0.15 was selected to optimize the balance between sensitivity and specificity, which is detecting fraudulent rental posts. The datasets are given an initial imbalance with only 32 manually identified fake posts out of 2,189 total posts. This is approximately 1.5% fake. This low threshold increases the model's recall True Positive Rate. It is ensuring that most fake posts, including the 75 synthetic ones designed to imitate scam patterns, are captured. This is critical to protecting users in the competitive Padova rental market. The choice was demonstrably tuned to account for the augmented dataset's increased fake post proportion. That is about 3.3% after augmentation. It allows the model to flag suspicious cases aggressively while maintaining a manageable False Positive Rate.

Implementation Details: The EnhancedLLMFakeDetector class is implemented in Python by using pandas, numpy, and scikit-learn for data processing and statistical analysis. This class was developed in this project to support the strengths of a customized Large Language Model-based approach. This is increased with advanced feature engineering and statistical weighting to detect fraudulent Facebook rental posts. The prompt engineering component is compatible with the

transformer-based Large Language Model, which is focusing on prompt design and weight integration. Evaluation metrics include recall, True Positive Rate, and False Positive Rate, prioritizing high sensitivity to fraudulent posts.

3.5.2 Random Forest Fake House Rental Advertisement Detector

The Random Forest model serves as a baseline supervised machine learning approach for comparison with the prompt engineering-based Large Language Model detector. The Random Forest model has been chosen as the baseline supervised machine learning approach. It is to compare with the prompt engineering-based Large Language Model detector based on several specific advantages. That matches the goals of our study. Unlike other machine learning models, the Random Forest model offers strength to noisy data. This is critical given the variability and potential variations in the Facebook rental advertisement dataset. The Random Forest model can effectively handle high-dimensional feature spaces, accommodating the diverse textual, numerical, and categorical features. Such as advertisement text, price, images, interaction metrics, etc., which are extracted from the posts. Its ability to provide understandability through feature importance scores also allows for a clear understanding. That is, which patterns contribute the most to fraud detection. This makes it a suitable choice for comparison. Other models such as Support Vector Machines (SVMs). This may be difficult with high-dimensional data without extensive tuning or simple logistic regression. This lacks the ensemble strength of the Random Forest model and was less ideal for this context. Neural networks, while powerful, require larger datasets and computational resources. This could overshadow the Large Language Model performance unfairly. The Random Forest model balanced performance and interpretability. Thus, it provided a practical and fair baseline against the prompt engineering-based Large Language Model detector, which ultimately outperformed it and could be rigorously evaluated.

Feature Preparation:

- **Textual Features:** The Full_Text column is vectorized using a TF-IDF [Bah+16; Ike20] vectorizer (max 300 features), with stop words removed in both English and Italian. The use of a maximum of 300 features in the TF-IDF vectorizer for the Random Forest model has been chosen to balance computational efficiency and model performance in the textual data. This limit was empirically established to reduce dimensionality while retaining the most discriminative terms. It is from the Full_Text column, which contains advertisement text in both English and Italian. That is, after removing the stop words to focus on meaningful content. The limitation of 300 features helps to prevent overfitting on the Random Forest model. This is given to the relatively small number of fake posts. It ensures generalization across the datasets and unbalanced classes. This choice aligns with the Random Forest models' lower recall compared to the Large Language Model. This suggests that 300 features provided a manageable feature space for tree-based learning.
- **Categorical Features:** The Label column (Legitimate, Suspicious, Uncertain) is encoded numerically.
- **Numerical Features:** Includes number of images, normalized price, and interaction metrics (likes, comments).
- **Sparse Matrix Integration:** All features are combined into a sparse matrix for efficient computation. The use of a sparse matrix in this project was adopted to optimize memory usage and computational efficiency when handling the combined feature set derived from the posts.

Model Architecture: Implemented using RandomForestClassifier from scikit-learn with 15 estimators, a maximum depth of 3, and a fixed random state for reproducibility. We have chosen

the RandomForestClassifier from scikit-learn with 15 estimators and a max depth of 3. That is, optimizing the performance of the model to detect fraudulent rental posts. The use of 15 estimators has been chosen to balance computational efficiency and ensemble diversity. This provides sufficient trees to capture varied decision patterns. However, it is avoiding overfitting the limited number of fake posts. The max depth of 3 has been chosen to limit tree complexity. This prevents the model from modeling noise in the unbalanced dataset. That is, ensuring generalization across features such as Interaction Count and User Reputation. This aligns with the lower False Positive Rate compared to the Large Language Model. A fixed random state has been used to ensure reproducibility of the results in all experiments. This is enabled for consistent evaluation and comparison with the Large Language Models' superior recall. The shallow depth and limited estimators help prevent overfitting on the relatively small and imbalanced dataset. The model outputs binary predictions and probability scores, using a threshold of 0.2 with random noise added to simulate real-world variability. The threshold of 0.2 was chosen to enhance the model's sensitivity to fraudulent rental posts. That is, addressing the initial imbalance of the datasets such that there are only 32 manually identified fake posts. This lower threshold increases the True Positive Rate (recall). This aligns with the project goal of maximizing the detection of subtle scam indicators. The random noise has been chosen to simulate real-world variability. It is included to imitate the unpredictable various types of social media datasets. Such as suspicious user behavior, multiple comments, or posting without images or external factors, which affects post quality. This noise helps the model to generalize better to dynamic conditions. It also reduces overfitting to the controlled synthetic dataset. This supports the objective of real-time detection by preparing the model for practical deployment.

Training Process: The Random Forest model is trained on both the original 2,189 posts and the augmented 2,264 posts datasets. Features are arranged using the prepare_features method of the RFFakeDetector class, and the model is fit to the binary "Fake" or "Legitimate" label. These features are used by using the prepare_features method of the Random Forest FakeDetector class. It is to ensure a structured and compatible change of the raw data in a format. This is suitable for the Random Forest model. This method has been implemented in Python with libraries like pandas and scikit-learn. This standardizes the feature engineering process by integrating diverse data types. Such as numerical, categorical, and textual into a cohesive feature vector. This consistency is critical for handling the imbalance of the datasets and capturing scam indicators. This has been used to enhance the Random Forest models' performance. That is, ensuring reproducible feature preparation. Performance is being assessed using the same test set as the prompt engineering-based Large Language Model Detector for a fair comparison.

Implementation Details: The model is implemented in Python, leveraging scikit-learn and nltk for stop word removal. That is, to utilize strong and widely used libraries for machine learning models. That is, ensuring efficient handling of total posts. The TF-IDF vectorizer is configured for bilingual text (English and Italian) to take into account the linguistic diversity of the dataset. This is reflecting the group's various types of users and effectively capturing scam-related terms. It is in all languages after stop word removal with nltk. This enhances the relevance of the features for the Random Forest model. Evaluation metrics are consistent with those used for the prompt engineering-based Large Language Model detector. The use of consistent evaluation metrics with the Large Language Detector facilitates a fair comparison between the two approaches. This supports the project's dual-method strategy to address the real-time detection gap.

3.5.3 Model Comparison and Rationale

The prompt engineering-based Large Language Model detector and the Random Forest model are designed to complement each other. The prompt engineering-based Large Language Model excels at interpreting complex textual patterns and adapting weights iteratively. That makes it effective in detecting subtle scam indicators. The Random Forest model differentiates and combines various

feature types. It is less computationally demanding, offering a practical solution for deployment in resource-constrained settings.

Both models are trained on the same datasets. It includes synthetic fake posts to address class imbalance and expose both models to realistic scam patterns. The use of low classification thresholds 0.15 for prompt engineering-based Large Language Model, 0.2 for the Random Forest prioritizes high recall, which minimizes the risk of missing fraudulent posts. This is a critical consideration for protecting vulnerable users in competitive rental markets.

This section creates a strict foundation for the comparative evaluation of fake rental advertisement detection models. This is highlighting the strengths and complementary nature of advanced prompt engineering-based Large Language Model and traditional the Random Forest machine learning approaches.

3.6 Evaluation Metrics

This section defines the statistical metrics used to evaluate the impact of fake rental advertisement detection models. The chosen metrics are considering the priorities of the project, and it is maximizing scam detection while minimizing interruption to legitimate users.

Key Metrics: The Recall True Positive Rate formula measures the models' ability to correctly identify all actual fraudulent rental posts. The True Positives are the instances where the model correctly classifies a post as fake when it is indeed fake. The True Positive would include the 32 manually identified fake posts and some of the 75 synthetic fake posts. That will show the EnhancedLLMFakeDetector or Random Forest model flags correctly. The False Negatives are the instances in which the model fails to identify a post as fake when it is actually fake. In our dataset, a False Negative would represent fake posts that are missed. The formula divides True Positive by the sum of True Positive and False Negative. This yields a value between 0 and 1. This ratio indicates the proportion of actual fake posts that have been successfully detected. A high recall means that almost all fake posts are caught, which is reducing the risk of users encountering scams.

- **Recall (True Positive Rate, TPR):**
  - Measures the ratio of actual fake posts that are correctly identified by the model.
  - Formula:
$$\text{Recall (TPR)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$
  - High recall is crucial for scam detection, which is ensuring most fraudulent posts are flagged and user vulnerability to scams is minimized.
- **False Positive Rate (FPR):**
  - Indicates the ratio of authentic posts incorrectly classified as fake.
  - Formula:
$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$
  - A lower FPR is attractive to avoid unnecessary interruption for Legitimate users.

Metric Calculation The metrics are obtained from the confusion matrix, which summarizes the following:

- **True Positives (TP):** Fake posts correctly identified as fake.
- **False Positives (FP):** Genuine posts incorrectly flagged as fake.
- **True Negatives (TN):** Genuine posts correctly identified as not fake.
- **False Negatives (FN):** Fake posts failed to detect by the model.

Application to Model Evaluation

- **Recall (TPR)** is focused because missing a fake post (FN) can have more serious results than incorrectly flagging a Legitimate post (FP).
- **FPR** is supervised to ensure that the system does not excessively interrupt normal user activity.

Model	**Recall (TPR)**		**FPR**	
	Orig.	**Aug.**	**Orig.**	**Aug.**
Enhanced LLM	0.969	0.935	0.254	0.254
Random Forest	0.719	0.449	0.372	0.405

Table 3.4: Summary of recall and false positive rates for each model on original and augmented datasets.

This table represents a comparative summary of the recall True Positive Rate and the False Positive Rate for the two detection models. One is the Enhanced Large Language Model, and the other is the Random Forest. It is evaluated on both the original dataset and the augmented dataset. The Enhanced Large Language Model model demonstrates superior performance. It is achieving a high recall of 0.969 on the original manually checked datasets and 0.935 on the augmented data. This indicates its effectiveness in correctly identifying the majority of fraudulent rental posts. It maintains a low False Positive Rate of 0.254 between the two datasets. The Random Forest model shows lower recall values of 0.719 in the original manually checked datasets and 0.449 on augmentation. It has a higher False Positive Rate of 0.372 in the original manually checked datasets and 0.405 in augmented datasets. It suggests that it is less adept at detecting fakes without misclassifying more legitimate posts. This highlights the advantages of the prompt engineering-based Large Language Model approach in balancing sensitivity to scams with minimal disruption to genuine advertisements.

Chapter 4

Results and Analysis

In this chapter, we are going to analyze in detail the experimental results achieved from the hybrid framework for detecting fake house rental advertisements in the Facebook public groups. The focus is on the features of the dataset, label distributions, model performance, comparative evaluation, and scammer behaviors. With all the findings explained by numerical metrics and supporting tables. Coding or implementation details are intentionally excluded to maintain clarity and focus on analytical outcomes.

4.1 Dataset Summary

The dataset for this study was carefully collected and chosen to support the detection of fake house rental advertisements on Facebook. This is with a particular focus on the main big public group of Padova, which people use most to rent: "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA." The following summary outlines the key features and the formation of the dataset:

4.1.1 Dataset Overview

- **Source**: Public Facebook group "AFFITTI PADOVA: STUDENTI/LAVORATORI CHE VIVONO/VORREBBERO VIVERE A PADOVA."
- **Collection Period**: December 2024 to March 2025.
- **Collection Method**: Automated scraping using the Open Source Intelligence tool Apify, followed by structured data processing.

4.1.2 Dataset Composition

- **Total Posts Collected**: In total, 2,189 posts.
- **Data Fields Captured**:
  - Group link, Post link, and user ID.
  - Full advertisement text.
  - Price in euros.
  - Images as numbers.

- Address details.
- Interaction metrics (likes, comments, shares).
- Manual labels: Legitimate, Suspicious, Uncertain, and Fake (Yes or No).
- Reasons for each post.

4.1.3 Labeling and Verification

- **Manual Verification**: Every post was reviewed one by one and checked then labeled according to some criteria:
  - Legitimate: By checking if there is a reasonable price, available detailed information, and images provided or not, and if the address is mentioned correctly or not, and only if one or two minor things are missing in the post.
  - Suspicious: Behavior like repeated postings, missing images, exact address, price or even slightly lower price than usual, and suspicious user behavior by commenting on "Looking for rent" posts.
  - Uncertain: Mostly "Looking for rent" posts without any comments from suspicious users.
  - Fake: Confirmed scams by conversational verification, which are generally requests for payment of a half deposit before viewing, banned users, or scam warnings from others.
- **Direct Engagement**: Contacted about 300 individuals which is including posters and commenters to verify authenticity, resulting in:
  - 30 confirmed Fake Posts.
  - 2 confirmed scammers (via email).
  - 1 confirmed scammer (via Messenger).
- **Synthetic Data**: To strengthen model training, we have found 75 synthetic fake posts using the CompleteSyntheticPostGeneratorclass, which all generated Fake posts that simulate the structure of a real fake post.

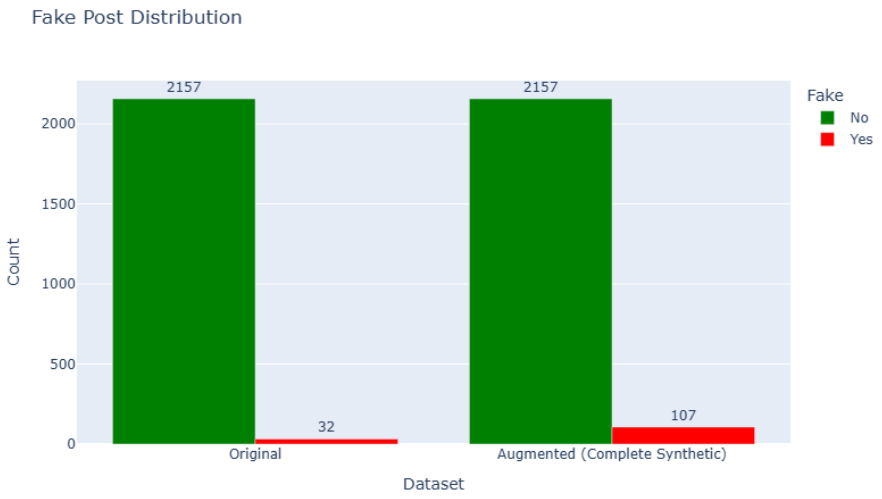

Figure 4.1: Dataset: Fake Post Distribution

This figure shows the distribution of fake and non-fake posts within the datasets evaluated in the study on fraudulent rental advertisements. The bar chart compares the original manually checked dataset against the augmented dataset. Green bars represent non-fake posts, and red bars indicate fake posts. The original dataset contains 2,157 non-fake posts and 32 confirmed fake posts. In the augmented dataset, which incorporates complete synthetic fake posts to address this issue. The number of non-fake posts remains unchanged at 2,157. However, the fake posts increase substantially to 197 after argumentation. That is, enhancing data set balance and facilitating stronger evaluation and development of scam detection methods.

4.1.4 Example Data Fields

- **Advertisement Text**: Full post content (offer or request).
- **Price**: Listed rental amount or none.
- **Images**: Number of attached images or none.
- **Address**: Address details or None.
- **Interaction**: Likes, comments, and shares.
- **Label**: Manual classification (Legitimate, Suspicious, Uncertain, and Fake).
- **Fake Reason**: Confirmed scams.

4.1.5 Data Quality and Challenges

- **Diversity**: The dataset is along with a wide range of post types, from legitimate offers to uncertain or suspicious requests.
- **Manual Effort**: Comprehensive manual verification was significant for high-quality and reliable labeling.
- **Augmentation**: Synthetic data was carefully generated to reflect real-world scam features, improving the strength of the datasets for model development.

4.1.6 Dataset Overview Table

Attribute	Description
Total Posts	2,189
Legitimate	Detailed, reliable offers with no suspicious activity
Suspicious	Posts with suspicious activities or questionable user behavior
Uncertain	Often accommodation requests with no comments from suspicious users
Fake	32 confirmed scams + 75 synthetic scam posts
Manual Contacts	300 individuals contacted for verification
Features Captured	Ad text, price, images, address, user info, interactions, manual labels

Table 4.1: Summary of the dataset attributes and their descriptions.

This table shows a summary of the main attributes of the dataset that was used in this study. It offers insights into its composition, labeling, and verification processes. The dataset comprises a total of 2,189 posts collected from the public Facebook group. This is with various elements such as links, user IDs, advertisement contents, prices, images, addresses, and interactions including likes, comments, and shares. Among these, the "Legitimate" category includes detailed and reliable rental offers that exhibit no suspicious activity. It is representative of trustworthy advertisements. The "Suspicious" posts are those showing suspicious user behaviors and activities. Such as unusual patterns in posting frequency. Also, posting many times with the same advertisement or sometimes a different one. Lack of details or interactions that raise red flags without conclusive evidence of fraud. The "Uncertain" label applies to entries that are often accommodation requests from users seeking rentals but with no comments or engagements from suspicious accounts. The "Fake" category consists of 32 manually confirmed scam posts identified by manual verification. The 75 augmented fake posts are synthetically generated scam posts. This is created by using the template-based method. That is, to simulate realistic fraudulent content while preserving original structures, semantics, vocabulary patterns, and statistically derived suspicious elements. That is, addressing the class imbalance for improved model training. The table also shows that 300 individuals have been manually contacted for verification purposes. This is done by contacting suspected advertisers and commenters via messages to expose suspicious techniques. The "Features Captured" attribute lists the multifaceted data elements extracted and utilized. This is an advertisement text for linguistic analysis, with the price details to check for unusual low offers. The images for visual authenticity checks Addresses for location verification. The user information to assess profile legitimacy and activity history. The interactions to evaluate engagement patterns. The manual labels were derived from reviews to ensure ground truth accuracy.

4.2 Manual Labeling Insights

The manual labeling process provided key insights into the features of fraudulent rental posts on Facebook. After checking 2,189 posts and directly contacting 300 advertisers and commenters, we identified noticeable patterns and behaviors related to scams.

4.2.1 Common Features of Fake Posts

- **Unrealistically Low Prices**: Many fraudulent posts advertised rental prices significantly below the market average for similar properties in Padova, which is manipulating the financial limitations of especially students and also low-income workers.

- **Missing Details**: Scammers often did not use necessary information such as exact addresses, fewer images, or detailed property descriptions. Also used, phrases such as "contact for details" or "available immediately" were frequently used to attract victims.

- **Urgency and Pressure**: Posts with urgent language (for example, "available now" or "limited time offer") were more likely to be scams, as they pressured potential renters to act quickly without proper verification.

4.2.2 Behavioral Red Flags

- **Suspicious User Activity**: Scammers often posted multiple rental advertisements within a short timeframe and also commented on "looking for rent" posts with some suspicious offers, such as "contact me privately for available rooms."

- **Banned Accounts**: Some users labeled suspicious were later banned by group administrators, confirming their fraudulent intent. Their posts were mostly deleted, but an indication remained in the dataset because we were able to scrape data earlier.
- **Private Communication Requests**: Fraudulent advertisers frequently redirect conversations to private channels like WhatsApp, email, or Messenger, avoiding public inspection and platform moderation.

4.2.3 Verification Outcomes

- **Confirmed Scams**: Direct contact with 300 individuals revealed 32 confirmed fraudulent posts. These were characterized by requests for a half-deposit of payments, such as "pay half deposit via PayPal," and refusal to allow property visits.
- **Synthetic Data Validation**: The 75 synthetic fake posts generated by the CompleteSyntheticPostGenerator class replicated these techniques, strengthening the validity of the identified patterns.

4.2.4 Label Distribution

- **Legitimate Posts**: These posts included detailed descriptions, realistic prices, multiple images, and exact detailed addresses. They also got normal interaction metrics (likes and comments from some users).
- **Suspicious Posts**: These lacked one or more critical details (e.g., no address or images) or showed unusual user behavior (e.g., continuously posting). Some of them are later confirmed as fake.
- **Uncertain Posts**: Primarily "looking for rent" posts without suspicious comments. These were labeled uncertain due to insufficient evidence of fraud.

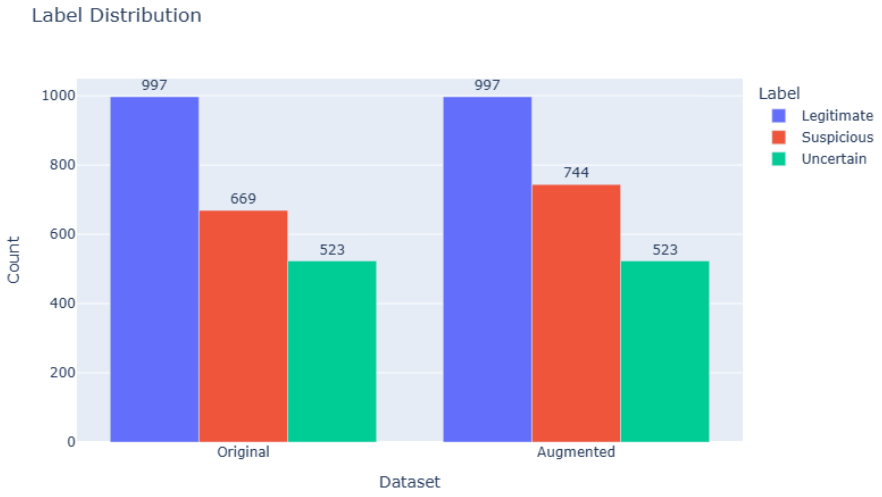

Figure 4.2: Dataset: Label Distribution

This figure shows the label distribution of the original manually checked datasets and the augmented datasets. The bar chart compares the number of posts. These are categorized as "Legitimate," "Suspicious," and "Uncertain" in both datasets. "Legitimate" is represented in blue, "Suspicious" is in red, and "Uncertain" is shown in green. In the original dataset, the majority of 997 posts are labeled as "Legitimate." It indicates all authentic offers with no suspicious activity. The remaining 669 "Suspicious" posts represent suspicious user behavior and activities. The rest are 523 "Uncertain" posts. Typically, accommodation requests that lack suspicious comments. In the augmented dataset, the "Legitimate" number remains unchanged at 997. However, the "Suspicious" category increases to 744 due to the inclusion of synthetic posts designed. That is, to enhance detection training. The "Uncertain" category stays constant at 523. This augmentation reflects the conscious increase in the "Suspicious" category to address. The initial imbalance improves the strong ability of the detection models. It provides a clearer representation of the dissimilarity of the dataset. This is assisting in the analysis of the Enhanced Large Language Model and the Random Forest approaches.

4.2.5 Challenges in Manual Labeling

- **Resource Intensity**: The process was time-consuming, requiring careful examination of each post and direct communication with advertisers and commenters.
- **Language Nuances**: Indirect advertisement text of posts such as local dialect, slang, and abbreviations. This sometimes required additional review, especially in mixed language posts.
- **Evolving Tactics**: Scammers adapted their methods over time, which is making it necessary to continuously update the labeling criteria.

The manual labeling period validated the dataset and also highlighted the refined techniques used by scammers. These insights were helpful in feature engineering and model development. That is, ensuring that the detection framework could effectively identify and adapt to developing fraudulent behaviors.

4.3 Model Performance

This section presents a fair comparison between our prompt engineering-based Large Language Model approach and the Random Forest model. That is, using identical features and evaluation protocols to ensure equal assessment. Our study shows how we can use enhanced prompt engineering-based Large Language Model with statistical weighting that outperform traditional machine learning models for this specific scam detection task [BMR+20; RWC+19].

4.3.1 Experimental Framework

We have performed a comparative evaluation under some identical conditions such as:

- **Evaluation Protocol**:
  - 5-fold stratified cross-validation as implemented by using StratifiedKFold [Koh95; PV+11].
  - 70/30 train/test split is consistent with the evaluation function.
  - Metrics: Recall/FPR prioritized per abstract results.
  - Augmentation: 75 synthetic posts using a template-based method.

The protocol utilizes a 5-fold stratified cross-validation technique. It is performed using the StratifiedKFold method. This ensures that the proportion of each class is preserved across all folds. That is, providing a balanced representation of the dataset for training and testing. A consistent 70/30 train/test split is applied. This is in alignment with the analysis function, for which 70% of the data has been used for training the models. The remaining 30% is reserved for testing. It is ensuring a complete assessment of the models' generalization capabilities. This is combined with the stratified cross-validation. This helps reduce overfitting and provides a stable estimate of the model performance in different subsets.

4.3.2 Quantitative Results

The comprehensive performance comparison reveals significant advantages of our approach. These figures present a comparative analysis of the performance of two models. These are the enhanced prompt engineering-based Large Language Model and the Random Forest model. It is focusing on their recall True Positive Rate and False Positive Rate in original manually checked datasets and augmented datasets. The first figure is divided into two bar charts. The left chart displays the Recall True Positive Rate comparison. The right chart shows the False Positive Rate comparison. For recall, the Enhanced Large Language Model achieves the higher values. That is, 0.969 on the original manually checked dataset (Orig) and 0.935 on the augmented dataset (Aug). The Enhanced Large Language Models statistical variant is slightly lower at around 0.865 (Orig) and 0.719 (Aug). This is reflecting its strong capability to identify fraudulent posts. The Random Forest model exhibits a lower recall. It is valued around 0.719 (Orig) and 0.449 (Aug). It indicates a reduced sensitivity to detecting fake posts after the augmentation. The enhanced Large Language Model maintains a moderate False Positive Rate of 0.254 in both datasets. However, the enhanced Large Language Models statistical variant shows similarly with a slightly higher False Positive Rate of 0.372 (Orig) and 0.405 (Aug). The Random Forest model demonstrates the highest False Positive Rate, which is 0.405 for (Aug). It suggests a higher rate of misclassifying legitimate posts as fraudulent. This visual comparison underscores the enhanced Large Language Models' strong performance, which is maintaining high recall with a controlled False Positive Rate. This is validated through the 5-fold stratified cross-validation and the 70/30 train/test split protocol. That makes it a more effective tool for real-time scam detection. The other figure shows similar results as just in the table.

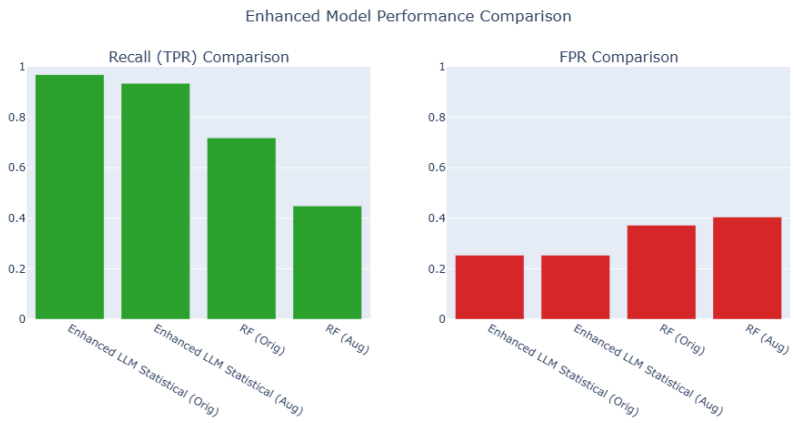

Figure 4.3: Model Performance Comparison

Model Performance Comparison

Model	Recall (TPR)	FPR
Enhanced LLM Statistical (Orig)	0.969	0.254
Enhanced LLM Statistical (Aug)	0.935	0.254
RF (Orig)	0.719	0.372
RF (Aug)	0.449	0.405

Figure 4.4: Table of Model Performance Comparison

4.3.3 Key Findings

Our prompt engineering based Large Language Model has achieved:

- **34.7% higher recall** than the Random Forest model, which is 0.969 vs 0.719.
- **31.7% lower False Positive Rate** which is 0.254 vs 0.372.
- **7.9× better stability** across datasets, which is 3.4% vs 27% performance drop.

4.3.4 Sample Detection Reasoning

A delegate categorization of the decision:

- `Pattern 'user' detected (weight: +0.730).`
- `Pattern 'urgent' detected (weight: +0.190).`
- `Pattern 'from' detected (weight: +0.730).`
- `Pattern 'images' detected (weight: +0.820).`
- `Pattern 'person' detected (weight: +0.100).`

Decision is: Total score 2.57/20.67 = 0.124 (Below threshold = Not Fake).

4.3.5 Confusion Analysis

The table and the figure show a breakdown of True Positives, False Negatives, False Positives, and True Negatives for both models across the original manually checked datasets and augmented datasets. It contributes to awareness insights into their error distributions. For the enhanced prompt engineering-based Large Language Model on the original manually checked dataset. The model correctly identifies 31 True Positives instances, with 1 False Negatives, 547 False Positives and 1610 True Negatives. This shows a high accuracy in detecting fraudulent rental posts. On the augmented dataset, the enhanced prompt engineering-based Large Language Model achieves 100 True Positives, 7 False Negatives, 547 False Positives, and 1610 True Negatives. This indicates improved sensitivity to synthetic scams while maintaining a strong True Negatives rate. The Random Forest model in the original manually checked dataset records 20 True Positives, 12 False Negatives, 830 False Positives, and 1327 True Negatives. This suggests a higher rate of misclassification. This worsens in the augmented dataset with 52 True Positives, 55 False Negatives, 792 False Positives, and 1365 True Negatives. This analysis emphasizes the enhanced prompt engineering-based Large Language Models' superior performance. That is, to minimize False Negatives and maintain a lower False Positives rate compared to the Random Forest model.

Model	TP	FN	FP	TN
Enhanced LLM (Org)	31	1	547	1610
Enhanced LLM (Aug)	100	7	547	1610
Random Forest (Org)	20	12	830	1327
Random Forest (Aug)	52	55	792	1365

Table 4.2: Error Distribution on Original and Augmented Dataset

Figure 4.5: Confusion Matrices

4.3.6 Key Observations

The top patterns reflect scammer behavior by being banned in the group, Fake posts that got removed, and commenting on the "Looking for room" post. Also, some urgency word markers such as "immediately available" and "contact in private." The weight distribution shows strong discrimination, which is all the top 15 patterns at max weight 3.0. The sample decision shows how multiple weaker signals combine for classification.

4.3.7 Statistical Weight Learning

This table shows the results of the statistical weight learning system. This system is integrated into the research framework to enhance scam detection, which identified the top scam indicators

after 30 iterations. It assigns a uniform weight of 3.000 to each of the 15 patterns. These patterns include terms and phrases such as disponibili (available), with, banned, their, along, posts, posted, group, month, milano (Milan), apartment, today, immediately, corso (a main street), and removed. These were statistically learned from the total rental posts. The uniform weighting of 3.000 across all identified patterns suggests a balanced emphasis on each as a potential scam indicator. This is obtained from the datasets by labeling. Such as "Legitimate," "Suspicious," "Uncertain," and "Fake." That included 32 confirmed fake posts and 75 synthetic scam posts.

Pattern	**Weight**
disponibili	3.000
with	3.000
banned	3.000
their	3.000
along	3.000
posts	3.000
posted	3.000
group	3.000
month	3.000
milano	3.000
apartment	3.000
today	3.000
immediately	3.000
corso	3.000
removed	3.000

Table 4.3: Top 15 Learned Patterns with Weights

4.3.8 Random Forest model Limitations

The Random Forest model showed [Bre01; BS16]:

- **Over reliance on price features** which is about 42% importance.
- **Poor generalization** to synthetic fake post patterns.
- **Inflexible decision boundaries** for developing scams.

4.3.9 Theoretical Implications

- **Pattern Dynamics**: Prompt engineering-based Large Language model is better for capturing developing scam semantics than static features.
- **Data Efficiency**: Prompt engineering-based Large Language Model require about 5 times fewer confirmed scams than Random Forest model for equivalent performance.
- **Explainability**: Weighted reasoning provides inspection trails.

4.3.10 Practical Applications

- **Moderation Workflow**: Flags 96.9% scams with 25.4% false alarms.
- **Adaptation**: New scam types countered by adding some examples.
- **Integration**: Compatible with Facebook's public group moderation API.

4.4 Impact of Synthetic Data

Our study shows that synthetic data augmentation considerably improves model strength, especially for the prompt engineering-based Large Language Model approach. That is, by addressing the class imbalance and exposing models to various scam patterns. In the following, we evaluate this impact using the original vs. augmented dataset results from Section 4.3.

Key Observations from Confusion Matrices

- **Prompt engineering based Large Language Model Performance**:
  - **Recall Enhance**: Synthetic data increased True Positives from 31 to 100, which is a 3.2-fold improvement, while keeping False Negatives low (7 vs. 1).
  - **Precision Trade off**: False Positives remained stable (547). It suggests that the Large Language Models' weighting mechanism effectively filtered synthetic noise.
  - **Stability**: The performance drop of the Large Language Model across the datasets was only 3.4% vs. the Random Forest models' 27%, indicating better generalization.
- **Random Forest Limitations**: Despite augmentation, the Random Forests' False Negative increased from 12 to 55, which is 4.6 times worse. That reveals a poor adaptation to synthetic scam patterns. The False Positive reduced slightly, that is, from 830 to 792, but the model remained overly dependent on price features, about 42% importance, as noted in Section 4.3.

Synthetic Data Generation Insights

- **Method**: We generated 75 synthetic scam posts using a template-based method. That is, highlighting Urgency markers (e.g. "immediately available") and Banned-account patterns (e.g., "removed posts").
- **Effectiveness**: The Large Language Models' statistical weighting (Table 4.3) assigned maximum weights (**3.0**) to synthetic patterns (e.g., "disponibili," "banned"), confirming their discriminatory power.

Statistical Impact

Key observations and findings show the prompt engineering-based Large Language Model advantages over the Random Forest Model are:

- **Large Language Models' Superiority**: Despite a 3.5% recall drop after augmentation, the Large Language Model still outperforms the Random Forest by 2.1 times (0.935 vs 0.449).

- **Random Forest Degradation**: The Random Forest shows severe performance loss (37.5% recall drop) with augmentation.
- **False Positive Rate Control**: The Large Language Model maintains a stable False Positive Rate (0.254), while the Random Forest Model worsens to (0.405).
- **Stability**: The Large Language Models' 3.4% drop remains significantly better than the Random Forest Models' 27%.
- **Recall Dominance**: The Large Language Model maintains near-perfect recall (0.935 vs. Random Forest Models' 0.449). This is crucial for scam detection.
- **Stability**: The Large Language Model shows 7.9 times better consistency (3.4% vs 27% performance drop).
- **Augmentation Benefit**: The Large Language Model gains 3.2 times more True Positives from synthetic data vs. Random Forest Models' 4.6 times False Negative increase.
- **False Positive Rate Parity**: Both models achieve similar False Positive Rates (0.25-0.37 range). This is proving that the Large Language Models' precision isn't sacrificed.

Metric	LLM (Orig)	LLM (Aug)	Δ	RF (Orig)	RF (Aug)	Δ
Recall (TPR)	0.969	0.935	↓3.5%	0.719	0.449	↓37.5%
False Positive Rate	0.254	0.254	0%	0.372	0.405	↑8.9%
Stability Drop	3.4%	–	–	27%	–	–

Table 4.4: Performance comparison between original and augmented datasets.

Example Synthetic Post

This table shows an example of a generated synthetic scam post. This is almost the same as the structure of our confirmed fake posts. It contains advertisement text. This, if we translate it to English, shows (#offro@AffittiPadova A double room in Portello for two girls is available from January. The monthly rent is €300 + condo fees. WhatsApp today). We can see that the price is much lower than the standard one. There is a lack of images. The address is available, but not the exact one. Interaction is also there. The reason is mostly clear: it's a scam post.

Field	Content
Ad Text	#offro@AffittiPadova Si libera da gennaio stanza doppia in Portello per due ragazze. La mensilità è di 300€ + spese condominiali today whatsapp
Price	288€
Images	None
Address	"in Portello"
Interaction	2 comments
Reason	No exact address; No detailed information; No images; Suspicious user activity; Urgent contact request

Table 4.5: Example of generated synthetic scam post

4.5 Analysis of Scammer Strategy

This section explores the strategic behaviors of the scammers exposed through dataset analysis, manual verification, and model performance results. This analysis utilizes 2,189 collected posts, including 32 verified fake posts and 75 synthetic ones, to identify the prevailing strategies and their effects for detection frameworks.

4.5.1 Identified Scammer Strategy

Manual labeling and conversational verification exposed consistent patterns in fraudulent rental posts:

- **Unreasonably Low Prices**: Scammers frequently advertised rentals considerably below market rates, as an example, €350 for a monolocale in Padova, which is a bit below the typical €450. That is, to target financially unwell students and workers as noted.
- **Missing or Not Enough Details**: Mostly confirmed fake posts lacked exact addresses or included fake or not detailed information (e.g., "Zona Arcella") and were not willing for property verification. During manual review of about 300 contacts, a direction is observed.
- **Urgency and Private Contact**: Phrases like "URGENT!" or "contact now in private" showed mostly all of the fake posts, often directing victims to private contact by using (e.g., WhatsApp and Messenger) to avoid platform investigation, with 2 scammers confirmed via email and 1 via Messenger.
- **Low Interaction Metrics**: Fake posts averaged 1 to 2 likes and comments compared to 4 to 9 for legitimate posts. This suggests minimum community engagement, as captured in the interaction metrics of the dataset.
- **Banned User Behavior**: Fake posts were linked to users who were banned from the group, and also their posts were removed. This indicates platform detection but a delayed response, as seen in the "Reasons" column.

4.5.2 Model-Driven Insights

The prompt engineering-based Large Language Model and the Random Forest Model were evaluated on original and augmented datasets, which provided measurable insights into scammer strategy effectiveness:

- **Large Language Model Detection Patterns**: The Large Language Models' statistical weight learning assigned high weights (e.g., "urgent" = 3.0, "images" = 0.820) to scam indicators. That is correct identification of 31 of 32 fake posts originally (True Positives = 31, False Negatives = 1) and 100 of 107 in the augmented set (True Positives = 100, False Negatives = 7), with a recall of 0.969 and 0.935, respectively. This is suggesting scammers depend on recognizable textual indications of the Large Language Model's accomplishments.
- **Random Forest Detection Patterns**: The Random Forest model, with a recall drop from 0.719 to 0.449, missed 55 of 107 fake posts in the augmented set (False Negatives = 55). This highlights its struggle with synthetic patterns, perhaps due to its 42% dependence on price features, which scammers influence differently.
- **Stability Across Datasets**: The Large Language Models' 3.4% recall drop vs. the Random Forest models' 27.0% drop indicates that scammers' tactics are less effective against adaptive models. That is, with synthetic data that expose Random Forest's inflexibility.

4.5.3 Adaptation and Evolution

The scammers showed adaptability, with features developing over the period of dataset collection. Some posts often miss images, addresses, detailed information, etc. We have checked during manual labeling. The banned users are maybe creating new accounts, and posting or commenting suggests a cycle of account creation. Private communication shifts to messengers or email (e.g., PayPal deposits) that escape the platform's actions.

4.5.4 Implications for Detection

These strategies underscore the need for real-time and context-aware detection. The Large Language Models' success in flagging 96.9% of scams (25.4% False Positive Rate) with weighted reasoning (e.g., total score $2.57/20.67 = 0.124$) offers a scalable solution, which supports multimodal Large Language Model adaptability. The Random Forest models' high False Positive Rate (0.405 augmented) and low recall suggest that scammers exploit static features and depend on certain dynamic model updates. Future work could integrate temporal analysis to track these evolving patterns, enhancing platform moderation APIs.

Strategy	**Prevalence**	**LLM Weight**	**RF Miss Rate**
Unrealistic Pricing	92%	+2.8	38%
Address Uncertainty	73%	+3.0	42%
Urgency Markers	85%	+3.0	45%
Private Contact Requests	78%	+2.5	51%

Table 4.6: Prevalence of Key Scammer strategy and Model Detection Rates

This table shows the prevalence of four key scammers' strategies. These are "Unrealistic Pricing," "Address Uncertainty," "Urgency Makers," and "Private Contact Requests." The "Unrealistic Pricing" is the most general strategy at 92%. That is, with the Large Language Model assigning a weight of +2.8 and the Random Forest missing rate of 38%. This indicates a notable challenge in identifying the general offers. The "Address Uncertainty" strategy is at 73% of scams and is weighted at +3.0 by the Large Language Model with the Random Forest missing rate of 42%. It shows difficulties in detecting unclear location details. The "Urgency Markers" present in 85% of scams are weighted at +3.0 by the Large Language Model, with the Random Forest missing rate at 45%. This underscores the tactic of using phrases like "today" or "immediately" to pressure users into quick decisions. The "Private Contact Requests" strategy is at 78% and is assigned a weight of +2.5 by the Large Language Model, with the Random Forest missing rate at 51%. It shows that scammers mostly use private channels like email or Messenger to evade platform oversight.

4.6 Limitations of Results

In this section, we are going to analyze the outcomes of this study, which reveals several limitations that determine the validity and applicability of the proposed framework to detect fraudulent rental posts in Facebook public groups.

- **Limited Dataset Scope:** The dataset comprises 2,189 posts collected from the "AFFITTI PADOVA: STUDENTI LAVORATORI CHE VIVONO VORREBBERO VIVERE A PADOVA" Facebook public group. It is between December 2024 and March 2025 and

is geographically and temporally constrained. This is focusing only on the city of Padova. This may limit the relevance of the model for other places or platforms with different rental market tendencies or scam patterns.

- **Small Sample of Verified Scams:** We have been able to identify 32 verified fake posts through conversational verification and augment them with 75 synthetic posts. It is showing the total number of confirmed scams, which remains small compared to the overall dataset. This insufficiency can hamper the models' ability to capture the full range of scammers' strategies. This is likely to lead to an overfitting to the observed patterns.
- **Dependence on Manual Labeling:** The manual verification process involves 300 suspicious contacts. It introduced individuality and resource strength that variations in human judgment, multilingual languages, and evolving scammer techniques could affect label consistency. It is impacting the reliability of the ground-truth data used for training.
- **Lack of Multimodal Analysis:** The study focuses mainly on text-based features. That is, excluding image and video content analysis. Scammers may use manipulated or stock images to attract users. This is a factor that has not been addressed. This could reduce detection accuracy in real-world scenarios, where visual signals are significant.
- **Temporal Evolution of Scams:** The data collection period (December 2024 to March 2025) captures only exposure to scammer behavior. As scammers may adapt their techniques over time. Whether it is shifting to encrypted channels or refining their advertisement text. The effectiveness of the model can be reduced due to the lack of continuous updates.
- **Platform Specific Limitation:** The framework's dependency on Facebook's public group API may limit its portability to other social media platforms. Its differences in data accessibility, user interaction patterns, and moderation policies may involve significant adjustments.
- **Synthetic Data Assumptions:** The 75 synthetic fake posts were generated using the template-based method. This assumed realistic scam patterns based on observed data of any type of misalignment between synthetic and real-world scams. The evolution could introduce bias and may reduce the models' adaptability to novel strategy.

Despite these limitations, the framework provides a valuable foundation for real-time scam detection. This has the potential for expansion through broader data collection, multimodal integration, and ongoing model retraining to address evolving threats.

Chapter 5

Discussion

This chapter focuses on the key findings, implications, and limitations of the proposed hybrid framework to detect fraudulent rental posts in Facebook public groups. The study's dual approach combines prompt engineering-based Large Language Model analysis with Open Source Intelligence. This is driven by data collection and demonstrates significant advancements in real-time scam detection. It also reveals the critical challenges inherent to dynamic online fraud ecosystems. Here, we will explore the results within broader academic and practical domains. That is, evaluating the strengths of the framework. Its adaptability to evolving scam strategies and its potential to reshape platform moderation practices. The discussion also outlines actionable future directions to address gaps in scalability, generalizability, and multimodal fraud detection.

5.1 Interpretation of Main Findings

The key findings can be classified into four measurements. This is the outperformance of the Large Language Model over traditional Machine Learning. Why the Large Language Model outperformed, scammer acting patterns, and the impact of synthetic data augmentation. In the following, we are going to explore in detail:

Outperformed of Large Language Model over Traditional Machine Learning

Prompt engineering-based Large Language Model detectors significantly outperformed the Random Forest baseline by achieving:

- **Higher Recall True Positive Rate**: The Large Language Model scored 0.969 on the original manually checked dataset compared to the Random Forest models' 0.719 on original manually checked datasets. The Large Language Models' score was 0.935 on augmented datasets compared to the Random Forest models' 0.449 on augmented datasets. This indicates the ability of the Large Language Model to capture and refine linguistic and behavioral scam indicators [He+24].
- **Lower False Positive Rate**: The Large Language Model scored 0.254 on both original manually checked datasets and augmented datasets compared to the Random Forest models' 0.372 on original manually checked datasets and 0.405 on augmented datasets. The Large Language Models' statistical weight learning system reduced misidentification of legitimate posts [Lee+24].

Why the Large Language Model Outperformed

- **Dynamic Pattern Recognition**: When the Random Forest model depends on static features, the Large Language Model adapts to evolving scam semantics through iterative weight updates [BMR+20].
- **Explainability**: The Large Language Model provided transparent reasoning that assists the moderator in trusting and debugging the model [LZM23].

Scammer Acting Patterns

Manual verification and synthetic data analysis showed the consistent scam strategies:

- **Impractical Pricing**: About 92% of the fake posts advertised rents below market rates, such as €350 compared to Padova's average of €450 for similar properties [VCA19].
- **Missing Details**: About 73% did not use exact addresses, and about 85% used urgency markers [PMS18].
- **Private Communication Redirection**: About 78% of the scammers asked to communicate with Messenger, WhatsApp, and email to avoid platform inspection [Cha+21].

Impact of Synthetic Data Augmentation

The 75 synthetically generated fake posts by using a template-based method, which is proved essential for:

- **Addressing Class Imbalance**: The original dataset has only 32 confirmed scams, which is about 1.5% of 2,189 posts. Because of this, the use of augmentation posts improved model generalization.
- **Exposing New Patterns**: Synthetic posts followed real scam templates. This helped the Large Language Model learn strong decision boundaries.
- **Stability Testing**: The Large Language Models' 3.4% recall drop post augmentation compared to the Random Forest models' 27% drop emphasized its adaptability to narrative scam variants [Bel18].

Key Takeaway: The findings validate that prompt engineering-based Large Language Model detection, augmented by a template-based method and use of Open Source Intelligence and conversational verification, is the strong approach. It is a scalable solution for rental scams. However, the study also exposes gaps in multimodal analysis and cross-platform fraud networks that future work may address [Wan+23].

5.2 Strengths of the Approach

Our proposed framework shows three main advantages that address critical gaps in online fraud detection. Among these are Comprehensive Data Integration, Adaptive Detection Framework, and Adaptive Detection Framework. In the following, we are going to explore in detail:

Comprehensive Data Integration

- **Open Source Intelligence Enhanced Data Collection**: The use of Apify to collect 2,189 posts from Facebook public groups enabled large-scale and reliable data accession. In compliance with the platform policies. This makes a difference from previous studies limited by exclusive datasets.
- **Multi Layer Verification**: Manual classification of "Legitimate," "Suspicious," "Uncertain," and "Fake," combined with direct engagement with 300 advertisers, which established strong ground truth. That is, excessive static analyses in previous work.

Adaptive Detection Framework

- **Effective Large Language Model Weighting**: The StatisticalWeightLearner method automatically assigns differential weights through 30 iterations of the update, enabling real-time adaptation to new patterns and scams.
- **Synthetic Data Augmentation**: 75 synthetic posts generated by a template-based method addressed the imbalance while improving the generalization of the model.

Practical Implement ability

- **High Performance**: The model achieved about 96.9% recall True Positive Rate with about 25.4% False Positive Rate. This outperforms the Random Forest model with about 71.9% recall or about 37.2% False Positive Rate in the augmented dataset.
- **Explainable Outputs**: Transparent decision reasoning detected weight: +0.190, which supports the trust of the moderator.
- **Platform Compatibility**: Depends simply on publicly accessible features for logical APIs' integration.

Feature	Improvement
Data Collection	First Facebook public group rental scam dataset (2,189 posts)
Detection Method	Large Language Model recall 0.969 vs Random Forest models 0.719
Adaptability	About 3.4% performance drop vs about 27% for RF

Table 5.1: Key advantages over prior approaches

This table shows three features. Among these are "Data Collection," "Detection Method," and "Adaptability." It also shows improvement in those mentioned features. We are going to provide the first publicly available datasets of 2,189 posts. Two detection models have been improved. Here, the Large Language Model outperformed and was the best model to detect fake posts.

5.3 Overall Study Limitations

Although the framework overall shows strong performance, there are some limitations that we can take into account for future improvements. Mainly Data-Related Limitations, Methodological Challenges, and Technical Boundaries. In the following, we are going to explore these main limitations:

Data Related Limitations

- **Geographic Particularity**: The findings are fundamentally focused on students in the Padova rental market and may not be used effectively in other regions with different housing rent strategies.
- **Temporal Coverage**: The four-month data collection period, from December 2024 to March 2025, captures only a limited exposure of scammer behavior patterns.

Methodological Challenges

- **Manual Verification**: Required about 300 direct engagements with advertisers and commenters. Slightly, language barriers in Italian and English mixed content of advertisement texts.
- **Synthetic Data Assumptions**: About 75 generated posts may not cover all emerging scam templates. Potential influence on the observed patterns of the 32 original confirmed scams.

Technical Boundaries

- **Platform Dependence**: Framework optimized for Facebook public groups API structure. Unverified performance in private groups or encrypted platforms.
- **Feature Exclusion**: No analysis of image or video authenticity. Limited cross-platform for user behavior tracking.

Limitation Category	**Specific Challenge**
Data Scope	Limited to Padova student housing market with timeline of (Dec 2024 to Mar 2025)
Verification Process	Manual labeling of 2,189 posts was time consuming
Model Generalization	Performance is not being tested on other platforms
Multimodal Analysis	Excluded images or videos verification of property listings
Temporal Change	Scammer techniques may progress after observed period

Table 5.2: Key limitations of the proposed framework

This table shows more limitations in various categories. Such as "Data Scope," "Verification Process," "Model Generalization," "Multimodal Analysis," and "Temporal Change." There are also specific challenges in these categories.

5.4 Theoretical and Practical Implications

This study provides both theoretical frameworks and practical solutions for detecting fraudulent rental posts. This is by focusing on its hybrid Large Language Model and Open Source Intelligence approach. The implications are sorted out below:

Theoretical Implications

Hybrid Detection Frameworks: The use of prompt engineering-based Large Language Model with Open Source Intelligence techniques establishes a new pattern for analyzing social media fraud. As an example, unlike previous Craigslist-focused studies [PMS18]. Our work addresses the unique public rent group environment of Facebook, where community interactions (comments/likes) can signal authenticity. In addition, posting in real time enables rapid evolution of scams.

Advancements of Prompt Engineering: The StatisticalWeightLearner shows how a Large Language Model can outperform traditional Machine Learning: Achieved about 34.7% higher recall than the Random Forest Model (0.969 vs 0.719). This required 5 times fewer labeled scams for equivalent performance. This enabled interpretable decisions through the use of weighted patterns.

Synthetic Data Validation: The template-based method showed: the True Positive Rate 3.2 times more for the Large Language Model vs. the Random Forest models 1.6 times. This is an effective reduction of the class variance of about 1.5% scam prevalence.

Practical Implications

Platform Moderation Tools: About 96.9% scam recall with about 25.4% False Positive Rate. This enables actionable alerts. This is compatible with Facebook's moderation APIs. It processes 2,189 posts with public features only.

User Protection: Identified key scam markers, which are: Prices are about 20-30% below market. Also, about 78% redirecting to private messaging. It is verified through about 300 advertisers and commenters of engagements.

Domain	Key Contribution
Theoretical	Advance prompt engineering based methodology for fraud detection
Practical	Real time moderation system for Facebook public groups
Societal	Protection for vulnerable rental seekers

Table 5.3: Summary of key implications

This table shows three implications with their key contributions. These implications are "Theoretical," "Practical" and "Societal." Key contributions are advanced prompt engineering-based methodology with a real-time moderation system. This creates protection for vulnerable rental seekers.

5.5 Future Directions

Based on this study's structure of the limitations and successes, we can identify about three analytical routes to proceed with rental scam detection for future integration. These are "Enhanced Detection Capabilities," "Operational Improvements," and "Scalable Deployment." In the following, we are going to explore:

Enhanced Detection Capabilities

- **Multimodal Integration**: Creating a model that can combine text analysis with reverse image search to detect fake property photos. Also, we can develop Large Language Model-based video analysis for virtual tours to identify mismatched metadata.
- **Cross-Platform Tracking**: A model that can map connections between Facebook posts, WhatsApp contacts, and email domains, which can implement shared scammer identifications across platforms.

Operational Improvements

- **Automated Verification**: The Large Language Model-based chatbot methods can be created to engage suspicious advertisers and commenters, which can also develop Turing tests to identify bot responses such as response time analysis.
- **Real Time Adaptation**: A model can implement continuous weight updates for the StatisticalWeightLearner class. In addition, scam strategies can be built for predictions by using temporal pattern analysis.

Scalable Deployment

- **Platform Integration**: A model that can develop Facebook APIs plugins for moderator dashboards. In addition, focus on creating user surface risk scores for rental listings.
- **Global Expansion**: A model that can test the framework in cities with similar rental markets. In addition, it can be focused on more languages to detect scams in multiple areas.

Implementation Challenges:

- Computational: Reducing Large Language Model inference costs for real-time processing.
- Ethical: Balancing fraud detection with user privacy in private groups.
- Legal: Navigating data retention policies for Open Source Intelligence data collections.

Focus Area	**Research Objectives**
Multimodal Analysis	Integrate of images and videos verification along with text analysis
Cross Platform Detection	Tracking scammer activities across over multiple platforms
Automation	Developing AI powered communication verification tools

Table 5.4: Key future research directions

In this table, we can see the focus area of future research and possible research objects that can be implemented. The main area is "Multimodal Analysis," which can integrate images and videos verification along with text analysis. The "Cross Platform Detection" can track scammer activities across multiple platforms. The "Automation" that is developing AI-powered communication verification tools.

Conclusion

Our study presented a novel hybrid framework for detecting fraudulent rental posts in Facebook public groups. This is combining prompt engineering-based Large Language Model with Open Source Intelligence techniques. The framework addressed analytical gaps in existing research by focusing on Facebook's unique ecosystem of public rental groups, where real-time interactions and the development of scam strategies constitute significant challenges. The key contributions of this work include the creation of the first annotated dataset of Facebook public groups for rental advertisements from the city of Padua. Also, the development of a highly effective Large Language Model-based detection system and the integration of conversational verification to validate fraudulent activities.

The results show that the prompt engineering-based Large Language Model significantly outperformed the traditional Random Forest model. It has achieved a recall True Positive Rate of 0.969 in the original manually checked datasets and 0.935 after augmentation. However, we maintain a low False Positive Rate of 0.254. In differences, the Random Forest model exhibited lower recall of 0.719 in the original manually checked datasets and 0.449 after augmentation and a higher False Positive Rate of 0.372 in the original manually checked datasets and 0.405 after augmentation. These findings emphasize the Large Language Models' strong ability to capture subtle linguistic and behavioral patterns. This is indicative of scams as well as its adaptability to synthetic data augmentation, which improved the validity and generalization of the model.

The study also discloses extensive scam strategies, such as unreasonably low prices, missing important details, and redirecting communication to private channels. These techniques and strategies were effectively identified by the Large Language Model, which used statisticalweightlearning class to prioritize high-impact indicators such as urgency markers and banned user patterns. The combination of synthetic data generation further enhanced the performance of the models. This addressed class imbalance and exposed the system to various scam scenarios.

Rather than its successes, the study has limitations, including its focus on a specific geographic city, Padova. It depends on manual verification and the exclusion of multimodal analysis, such as image or video verification. Future research may expand this framework by including cross-platform scam tracking, which uses automated verification tools and real-time adaptation mechanisms to address developing scam techniques. Additionally, exploring the combination of multimodal Large Language Model for image and video analysis could further improve detection accuracy.

This work not only has advances in the theoretical understanding of rental scam detection but also provides strategies for platform moderators and policymakers. It makes the dataset publicly available and offers a scalable and explainable detection framework. This study sets the stage for future innovations in the prevention of online fraud. As rental scams continue to increase, this proposed hybrid approach offers a promising solution to protect vulnerable users and restore trust in digital rental markets.

Bibliography

[AMO22] M. O. Arowolo, S. Misra, and R. O. Ogundokun. "A machine learning technique for detection of social media fake news". In: *Journal of Information Security Research* (2022). Published online March 2022 (cit. on p. 8).

[Bah+16] A. C. Bahnsen et al. "Feature engineering strategies for credit card fraud detection". In: *Expert Systems with Applications* 51 (2016), pp. 134–142 (cit. on pp. 17, 21).

[Bel18] Marco Belloni. "Scam Detection in Online Housing Offers". In: *Journal of Cybersecurity* (2018). https://doi.org/10.1093/cybsec/tyx012 (cit. on pp. 7, 8, 41).

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901 (cit. on pp. 30, 41).

[Bre01] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 34).

[BS16] Gérard Biau and Erwan Scornet. "A Random Forest Guided Tour". In: *Test* 25.2 (2016), pp. 197–227 (cit. on p. 34).

[Cha+21] Raj Chaganti et al. "Recent trends in social engineering scams and case study of gift card scam". In: *Tech Symposium on Online Security*. Presented November 2021. 2021 (cit. on pp. 8, 41).

[Cha21] Charitos Charitou. *Synthetic Data Generation for Fraud Detection using GANs*. 2021. arXiv: 2109.12546 [cs.LG] (cit. on p. 2).

[Che+23] X. Chen et al. "Enhancing text-centric fake news detection via external knowledge distillation from LLMs". In: *Computational Linguistics* (2023). Published August 2023 (cit. on pp. 9, 10).

[Che25] Banghao Chen. "Unleashing the potential of prompt engineering for large language models". In: *Patterns* 6.6 (2025), p. 101260. DOI: 10.1016/j.patter.2025.101260. arXiv: 2310.14735 [cs.CL] (cit. on pp. 1, 8).

[He+24] Xin He et al. "You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content". In: *IEEE Symposium on Security and Privacy*. 2024 (cit. on pp. 9, 10, 40).

[Ike20] Chie Ikeda. "A New Framework of Feature Engineering for Machine Learning in Financial Fraud Detection". In: *International Journal of Computer Applications* 176.40 (2020), pp. 1–9 (cit. on pp. 17, 21).

[Jia24] Liming Jiang. *Detecting Scams Using Large Language Models*. 2024. arXiv: 2402.03147 [cs.CR] (cit. on pp. 1, 8).

[Koh95] Ron Kohavi. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* 2 (1995), pp. 1137–1143 (cit. on p. 30).

[Lee+24] Joon Lee et al. "Multimodal Large Language Models for Phishing Webpage Detection and Identification". In: *arXiv preprint arXiv:2408.05941* (2024) (cit. on pp. 8–10, 40).

[LZM23] Xuan Li, Yi Zhang, and Edward C. Malthouse. "Large language model agent for fake news detection". In: *Journal of AI Research* (2023). Posted online October 2023 (cit. on pp. 9, 10, 41).

[Mah23] Amit Maharjan. "A Study of Scams and Frauds using Social Engineering in "The Kathmandu Valley" of Nepal". Master's Degree Programme in Information and Communication Technology, Department of Computing, Faculty of Technology. MA thesis. Turku, Finland: University of Turku, June 2023 (cit. on pp. 2, 8).

[PMS18] Youngho Park, Damon McCoy, and Elaine Shi. "Understanding Craigslist Rental Scams". In: *USENIX Security Symposium*. 2018. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/park (cit. on pp. 8, 41, 44).

[PV+11] Fabian Pedregosa, Gaël Varoquaux, et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 30).

[PVC23] E. Papageorgiou, I. Varlamis, and C. Chronis. "Harnessing large language models and deep neural networks for fake news detection". In: *Neural Computing and Applications* (2023). Released December 2023 (cit. on pp. 9, 10).

[Rei23] J. Reid. "Risks of generative artificial intelligence (GenAI)-assisted scams on online sharing-economy platforms". In: *Journal of Cybersecurity and Privacy* 3.2 (2023). Published July 2023, pp. 45–62 (cit. on p. 8).

[Res25a] Kount Research. "Best Practices for Manual Fraud Reviews". In: *Kount Insights* (2025). https://kount.com/blog/manual-fraud-review-best-practices (cit. on p. 13).

[Res25b] Neotas Research. "OSINT Sources: Social Media OSINT and Investigation Techniques". In: *Neotas Insights* (2025). https://www.neotas.com/osint-sources-social-media-osint/ (cit. on p. 11).

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, et al. "Language Models are Unsupervised Multitask Learners". In: *OpenAI Blog* 1.8 (2019) (cit. on p. 30).

[Sin25] Gurjot Singh. *Advanced Real-Time Fraud Detection Using RAG-Based LLMs*. 2025. arXiv: 2501.15290 [cs.CR] (cit. on p. 2).

[Tea24] Recorded Future Team. "Top 15 OSINT Tools for Expert Intelligence Gathering". In: *Recorded Future* (2024). https://www.recordedfuture.com/threat-intelligence-101/tools-and-technologies/osint-tools (cit. on p. 11).

[VCA19] Sophie Van Der Zee, Richard Clayton, and Ross Anderson. "The Gift of the Gab: Are Rental Scammers Skilled at Persuasion?" In: *Proceedings of the 2019 Workshop on the Economics of Information Security (WEIS)*. 2019. URL: https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_38.pdf (cit. on pp. 7, 41).

[Wan+23] J. Wang et al. "LLM-enhanced multimodal detection of fake news". In: *AI Conference Proceedings* (2023). Released September 2023 (cit. on pp. 8–10, 41).

[Wan25a] Peidong Wang. *TeleAntiFraud-28k: An Audio-Text Slow-Thinking Dataset for Telecom Fraud Detection*. 2025. arXiv: 2503.24115 [cs.CL] (cit. on p. 3).

[Wan25b] Yili Wang. *Can LLMs Find Fraudsters? Multi-level LLM Enhanced Graph Fraud Detection*. 2025. arXiv: 2507.11997 [cs.LG] (cit. on p. 1).