MOZ-Smishing: A Benchmark Dataset for Detecting Mobile Money Frauds

Felermino D. M. A. Ali[1,2,3], **Saide M. Saide**[3], **Rui Sousa-Silva**[2], **Henrique Lopes Cardoso**[1]

[1]LIACC, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

[2]CLUP, Faculdade de Letras da Universidade do Porto, Via Panorâmica, 4150-564, Porto, Portugal

[3]DEI, Faculdade de Engenharia da Universidade Lúrio, Pemba, 3203, Cabo-Delgado, Mozambique

{up202100778, hlc}@fe.up.pt, saide.saide@unilurio.ac.mz, rssilva@letras.up.pt

Abstract

Despite the increasing prevalence of smishing attacks targeting Mobile Money Transfer systems, there is a notable lack of publicly available SMS phishing datasets in this domain. This study seeks to address this gap by creating a specialized dataset designed to detect smishing attacks aimed at Mobile Money Transfer users. The data set consists of crowd-sourced text messages from Mozambican mobile users, meticulously annotated into two categories: legitimate messages and smishing attempts. The messages are written in Portuguese, often incorporating microtext styles and linguistic nuances unique to the Mozambican context. We also investigate the effectiveness of LLMs in detecting smishing. Using in-context learning approaches, we evaluate the models' ability to identify smishing attempts without requiring extensive task-specific training. The data set is released under an open license at the following link: https://huggingface.co/datasets/MOZNLP/MOZ-Smishing

1 Introduction

Mobile Money Transfer (MMT) systems have emerged as a transformative financial technology, particularly in developing countries where traditional banking infrastructure is often inadequate or inaccessible. These systems have revolutionized financial inclusion by providing essential services to underserved populations, enabling users to deposit, withdraw, transfer money, pay for goods and services, and access credit and savings—all through the convenience of a mobile device. In regions such as Sub-Saharan Africa, where traditional banking adoption remains low, MMT systems have become a cornerstone of economic activity and financial empowerment.

According to GSMA (2024b), the global adoption of MMT systems has reached unprecedented levels, with over 1.75 billion registered accounts worldwide as of 2024. These systems process an estimated $1.4 trillion annually, equivalent to approximately $2.7 million per minute. Sub-Saharan Africa has emerged as the most active region for MMT adoption, driven by the widespread use of platforms such as M-Pesa, Airtel Money, and MTN Mobile Money. However, this rapid growth has also attracted the attention of cybercriminals, making MMT users increasingly vulnerable to fraud (INTERPOL, 2020).

Mobile money fraud has become a significant concern across Africa, with the number of victims rising sharply in recent years. This alarming trend underscores the urgent need for fraud detection and mitigation strategies. Therefore, various solutions have been proposed to address this challenge (GSMA, 2024a), with a growing emphasis on leveraging advanced technologies such as Artificial Intelligence and Machine Learning to detect and prevent fraudulent activities (Delvia Arifin et al., 2016; Balim and Gunal, 2019; Ghourabi et al., 2020; Ghourabi, 2021; Jain and Gupta, 2018, 2019; Jain et al., 2020; Mishra and Soni, 2020, 2021; Roy et al., 2020; Sonowal and Kuppusamy, 2018). However, the scarcity of high-quality, domain-specific datasets hinders the development of effective AI-based fraud detection systems. These solutions are inherently data-hungry, requiring a large amount of labeled data to train deployable models. Unfortunately, few publicly available datasets exist for smishing identification and other types of mobile money fraud, limiting the progress of research in this critical area.

In this study, we aim to bridge this gap by contributing a benchmark dataset specifically designed for smishing identification in the context of MMT. This dataset is constructed to reflect real-world scenarios and includes a set of smishing attempts targeting real mobile money users. Additionally, we evaluate the performance of existing LLMs using in-context learning techniques to assess their effectiveness in detecting smishing attempts. Our

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 158–166

findings provide valuable insights into the potential of LLMs for fraud detection and highlight areas for future research and development.

2 Literature Review

One of the most widely used datasets for smishing detection is the one proposed by Almeida et al. (2013). This dataset contains 5,574 text messages, divided into 4,827 legitimate messages and 747 fraudulent messages. While this dataset has been influential in advancing research in smishing detection, it has notable limitations. First, the dataset contains a relatively small number of smishing examples, which may limit the generalizability of models trained on it. Second, the dataset is exclusively composed of English-language text messages, which restricts its applicability to non-English-speaking regions where smishing fraud is also prevalent.

Other publicly available datasets, such as those proposed by Timko and Rahman (2024) and Chen and Kan (2012), also focus primarily on English-language content and general smishing or spam messages, rather than targeting the specific context of mobile money fraud. While these datasets have contributed to the development of spam and fraud detection systems, they do not adequately address the unique linguistic and contextual nuances of MMT-related fraud, particularly in regions where English is not the primary language.

To address the language gap, some researchers have proposed datasets that include other non-English languages. For example, Yadav et al. (2011), Ghourabi (2021) and Mambina et al. (2022), have developed datasets that besides English also included Hindi, Arabic and Swahili respectively.

In general, all existing data sets often lack a specific focus on mobile money fraud, instead addressing more general forms of SMS spam or smishing. Our work seeks to address these gaps by introducing a novel dataset focused on Portuguese-language text messages, with a particular emphasis on smishing attempts targeting MMT users. Similar to Mambina et al. (2022); Timko and Rahman (2024), this data set was constructed using community-based approaches, where we crowd-sourced both smishing and legitimate messages.

3 Dataset Collection

We gathered data from users of MMT services in Mozambique, a country currently experiencing a wave in the adoption of such services. The MMT landscape in Mozambique is dominated by several prominent platforms, including M-PESA, E-Mola, and mKesh, which are operated by the country's major telecom providers: Vodacom, Movitel, and Tmcel, respectively. However, the rapid growth of these services has also led to an increase in fraudulent activities targeting users. For instance, Vodacom, the operator of M-Pesa, reported that approximately 80 people fall victim to fraudulent mobile money transactions daily in Mozambique. This alarming trend underscored the necessity to study and understand these scams. To address this, we crowd-sourced fraudulent messages from users, including those who had already been victimized by such schemes.

The data collection methodology comprised the following steps:

Crowdsourcing Smishing Messages: We launched a campaign inviting people to join a dedicated WhatsApp group. Participants were encouraged to share suspicious or fraudulent text messages they had received, particularly those from unknown sources that appeared to target their mobile money accounts. Clear instructions were provided to guide participants in identifying these messages, emphasizing the importance of sharing only those texts that they believed were attempts to defraud them or cause financial loss. Participants could share these messages either by submitting screenshots or forwarding the text directly to the group.

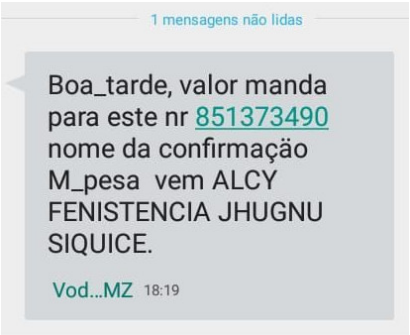

Figure 1: A Sample of a smishing text message.

Crowdsourcing Legitimate Messaging: Similarly, we invited participants to share messages that they considered legitimate. We encouraged them

to submit messages related to MMT topics, as well as other non-fraudulent messages. This helped us build a balanced data set for comparative analysis.

Data Preprocessing: We preprocessed the collected data by performing the following steps. First, all message screenshots were transcribed in plain text format. Next, we identified and removed duplicate messages. Finally, all personal identifiers within the legitimate messages were anonymized to ensure user privacy.

The final dataset contains 552 instances of smishing messages and 2,009 legitimate text messages. Figure 2 illustrates the embedding space of both categories using UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) clustering (McInnes et al., 2020), highlighting their distribution. Furthermore, Table 1 presents a sample of 8 data points, which showcases examples of legitimate and smishing messages from our dataset.

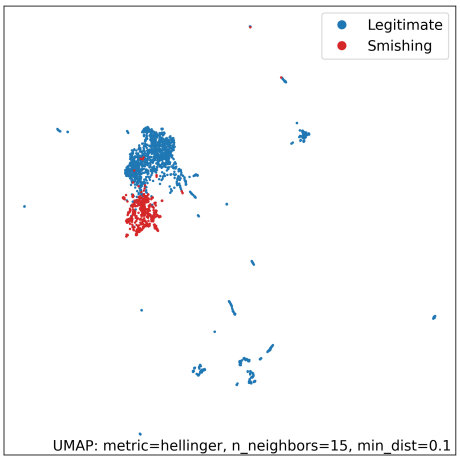

Figure 2: UMAP clustering, where blue points represent legitimate messages, whereas red points are smishing messages

4 Exploratory Data Analysis

4.1 Smishing Tactics

To further our analysis of the tactics used by scammers, we conducted a content analysis on the collected smishing messages. Our analysis identified several recurring patterns and social engineering tactics used by scammers. These tactics primarily aim to deceive users into transferring monetary funds directly or inadvertently, ultimately resulting in financial loss. We identified the following tactics:

Bulk SMS: We collected a total of 692 text messages from our dataset. After preprocessing, we identified that 140 messages were duplicates. Interestingly, the persistence of duplicate messages provided valuable information on the operational strategies of scammers. Since identical messages appear to be disseminated to a large number of recipients via multiple phone numbers, it suggests that scammers target various random recipients simultaneously, thereby increasing the chances that at least some victims will fall into their trap. Furthermore, we observed that scammers frequently used different accounts or contact numbers in various messages. This deliberate strategy presumably serves as a mechanism to avoid detection and tracking.

Pretending an Existing Transaction was Previously Arranged: Scammers create a psychological trap that a transaction was previously agreed upon by vaguely referencing prior interactions or conversations, as exemplified by ambiguous phrases like:

- *"aquele valor"* ("that amount of money").

Creating Urgency and Pressure: Scammers attempt to induce panic or urgency, prompting immediate action from their victims. Typical tactics used by scammers manipulate victims into quick, and often irrational, include using the following phases:

- *"manda agora"* ("send now");
- *"tem problema a minha conta M-pesa"* ("my M-pesa account has a problem");
- *"meu telefone caiu em água"* ("my phone fell into water");
- *"já podes mandar"* ("you can send it now").

Impersonation of Trusted Parties or Familiar Contacts: Scammers use impersonation techniques that involve pretending to be trusted persons such as family members or friends. They frequently use informal language and familiar salutations such as *"amigo/a"* ("friend"), *"man"*, or typical greetings such as:

- *"oi," "boa tarde," "bom dia"* (informal salutations).

Text Message	Target Label
Bom dia pai sou eu sua filha estou a espera desse valor quero pagar matricula	Legitimate
Bom dia bro, podes mandar aquele valor para o meu número aguardo teu sinal	Legitimate
Bom dia Rosinha peço para me mandar 500 Mts no M-Pesa pago no final do mês	Legitimate
Kmk brow, tudo bem? Peço que me envies aquele valor para minha conta m-pesa, estou a precisar.	Legitimate
Manda o valor neste número,858773567.M-pesa vem em nome de Manuel Vasco R.Ok	Smishing
bom dia, este valor enviame nesta conta: 857491433 vem em nome de ROSA MILIONE FERRO	Smishing
Esta bem.O valor podes mandar para este Nr 841898297 vem em nome e Castro Jos Fabio!	Smishing
Man Esse Valor Manda Neste Numero 857170842 M,pesa Vem Abel Vasco	Smishing

Table 1: Sample messages from the dataset. Phone numbers used to receive fraudulent payments are shown in blue, the MMT platforms exploited by scammers are marked in red, and the names under which fraudsters registered their MMT accounts are highlighted in green.

Impersonating Common Names: Scammers increase the authenticity and credibility of scam messages by carefully selecting common local names. The names identified in the messages include:

- Top 5 frequent **First Names:** *"Maria", "Luisa", "Alberto", "Ana".*
- Top 5 frequent **Surnames:** *"João", "José", "Mário", "Joaquim", "Manuel"*
- **Mozambican family names:** *"Siquice", "Chacuanda", "Nhampossa", "Páisse", "Mustafa", "Mapisse", "Nhalungo", "Cuamba", "Mutucua", "Machava", "Malangisse", etc.*

Fake Technical or Emergency Problems: Many messages exploit scenarios involving fictitious technical difficulties or emergencies to justify the use of an unfamiliar phone number. Frequent examples found in messages are:

- *"minha conta tem problema"* ("my account has an issue"),
- *"meu número não tá receber dinheiro"* ("my number can't receive money anymore"),
- *"telefone desligado," "telefone caiu na água"* ("phone is off," "phone fell in water").

Politeness and False Courtesy: Scammers strategically incorporate polite and courteous expressions into their messages, lowering the victims' guard and diminishing suspicion. Instances include phrases such as:

- *"desculpe pelo incómodo"* ("sorry for the inconvenience"),
- *"por favor"* ("please"),
- *"bom dia," "boa tarde"* ("good morning," "good afternoon").

Small Mistakes, Microtext, and Typographical Errors: Finally, deliberate typographical errors or microtext were frequently observed in smishing messages, making them resemble authentic informal texts. We noticed many intentionally casual errors or informal grammar, thus giving messages a natural, rushed appearance. Scammers may also use these errors to avoid automated filtering or spam detection systems. Examples include abbreviations, improper capitalization, simplified spelling, or grammatically inconsistent phrases, making the messages appear realistic and spontaneous, and reducing skepticism.

4.2 Mobile Money Platforms used by Scammers

Our analysis revealed that scammers frequently exploit various MMT platforms to receive illicit funds. Among the most commonly used platforms are M-Pesa, E-mola, and Ponto-24. We observed a strong preference for the use of M-Pesa. This preference may be attributed to M-Pesa's status as one of the oldest and largest MMT platforms in the market, with a widespread user base and high transaction volumes. However, this trend also highlights a critical vulnerability within these platforms, as they appear to be susceptible to exploitation by criminals for this type of illicit activity. The lack of robust mechanisms to track and flag suspicious transactions on these platforms further exacerbates the problem.

Furthermore, our analysis revealed that the phone numbers used to receive fraudulent funds are typically unique and not reused in different smishing messages (see Figure 3). This suggests that scammers use a "one-time use" strategy for these numbers, likely to avoid detection and complicate efforts to trace the transactions. Interestingly, we identified a recurring pattern in the phone

numbers used by these criminals. Specifically, the numbers often featured consecutive prefixes (see Figure 4), indicating that attackers may have access to a sequence of SIM cards purchased in bulk. This pattern implies a level of organization and resourcefulness among the scammers, as they appear to systematically acquire and deploy multiple SIM cards to facilitate their schemes.

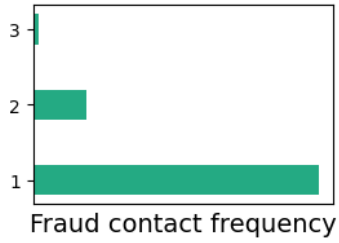

Figure 3: Phone number frequency on smishing messages

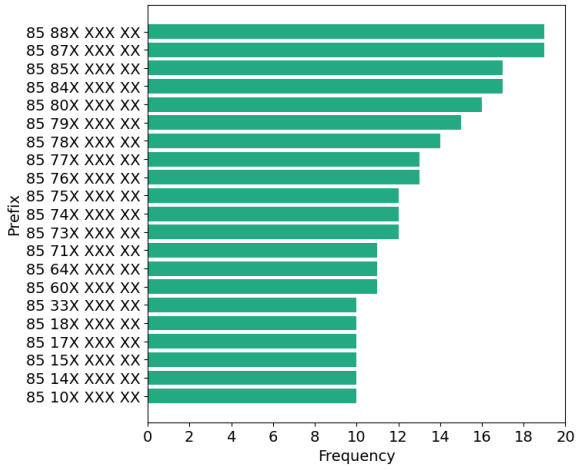

Figure 4: Top frequent four digits prefix

5 Experiments and Results

This section describes our experimental setup, presents the results from benchmarking various LLMs for smishing detection, and discusses the implications of these results in the context of mobile money transfer fraud detection. Specifically, we explore in-context learning capabilities across multiple LLMs using various few-shot prompting scenarios.

5.1 Experimental Setup

Using our newly constructed dataset, we conducted experiments to evaluate the effectiveness of state-of-the-art LLMs in detecting smishing messages. The selected models for our evaluation included *Dolly-v2-12B* (Conover et al., 2023), an open-source conversational model developed by Databricks; *Mistral-Small-24B* (Jiang et al., 2024), developed by Mistral AI; *Qwen2.5-14B*, a multilingual language model introduced by Alibaba (Yang et al., 2024); and *EuroLLM-9B*, an LLM specially optimized for multilingual European language tasks (Martins et al., 2025).

Each model was assessed using an in-context learning approach, in which carefully designed prompts incorporated balanced examples of legitimate and smishing messages. Furthermore, model performance was evaluated under multiple learning scenarios, including 0-shot and few-shot settings. To ensure consistency and reproducibility, all models received a standardized prompt (see Figure 5 and Figure 7), outlining the task and providing examples labeled as "*Legitimate*" or "*Smishing*".

Prompt Template

Below are examples of messages classified as Positive (indicating intent of smishing or phishing) or Negative (indicating no intent of smishing or phishing):

Input: "*Bom dia, o valor melhor mandar para este nr 858798603 Mpesa nome Israel Robate Charimba, o meu atingiu limite.*"
Output: Positive

Input: "*Irmã peço pra me mandar mil mt.*"
Output: Negative

...

Input: "*Ok Aquele Valor Manda Para Este Nr D M-pesa 846861650 E Nome D Essinate Jofres*"
Output:

Figure 5: Example of the few-shot prompt template

We measured the performance of each model using commonly adopted evaluation metrics in binary classification tasks, including the F1-score for each class (Smishing and Legitimate), and the Macro-F1 average across the classes to account for potential imbalances in class distribution.

All experiments were executed on 4 NVIDIA A10 GPU cards.

5.2 Experimental Results

The results of our experiments are presented in Table 2. Qwen2.5-14B notably achieved the highest overall performance among the evaluated mod-

#shot	F1	EuroLLM-9B pt	EuroLLM-9B en	dolly-v2-12b pt	dolly-v2-12b en	Qwen2.5-14B pt	Qwen2.5-14B en	Mistral-Small-24B pt	Mistral-Small-24B en
0-shot	Legitimate	0.69	0.59	0.45	0.48	**0.69**	**0.63**	0.51	0.51
	Smishing	0.43	0.41	0.3	0.28	**0.49**	**0.48**	0.45	0.45
	Macro	0.56	0.5	0.38	0.38	**0.59**	**0.55**	0.48	0.48
1-shot	Legitimate	0.0	0.54	0.2	0.51	**0.63**	0.56	0.54	**0.63**
	Smishing	0.33	0.42	0.33	0.32	**0.5**	0.47	0.46	**0.5**
	Macro	0.17	0.48	0.26	0.41	**0.56**	0.51	0.5	**0.56**
2-shot	Legitimate	0.4	0.34	0.15	0.59	**0.62**	0.6	0.58	**0.65**
	Smishing	0.4	0.41	0.37	0.39	**0.5**	0.49	0.48	**0.52**
	Macro	0.4	0.38	0.26	0.49	**0.56**	0.55	0.53	**0.58**
4-shot	Legitimate	0.22	0.52	0.08	0.65	**0.7**	**0.72**	0.62	0.69
	Smishing	0.36	0.46	0.36	0.42	**0.54**	**0.56**	0.5	0.54
	Macro	0.29	0.49	0.22	0.53	**0.62**	**0.64**	0.56	0.62
6-shot	Legitimate	0.32	0.56	0.05	0.64	**0.75**	**0.77**	0.68	0.74
	Smishing	0.38	0.47	0.36	0.42	**0.58**	**0.6**	0.53	0.57
	Macro	0.35	0.51	0.21	0.53	**0.67**	**0.68**	0.61	0.65
8-shot	Legitimate	0.32	0.62	0.03	0.71	**0.79**	**0.8**	0.72	0.78
	Smishing	0.38	0.5	0.36	0.46	**0.61**	**0.62**	0.55	0.6
	Macro	0.35	0.56	0.19	0.59	**0.7**	**0.71**	0.63	0.69
16-shot	Legitimate	0.67	0.8	0.08	0.87	**0.87**	**0.86**	0.78	0.83
	Smishing	0.5	0.62	0.36	0.53	**0.71**	**0.69**	0.6	0.65
	Macro	0.58	0.71	0.22	0.7	**0.79**	**0.78**	0.69	0.74

Table 2: Performance of the models different few-shot settings with Portuguese and English prompts, with the highest values shown in bold.

els, with F1-scores consistently higher in most scenarios, reaching a Macro F1-score of 0.79 in the 16-shot learning setting. Mistral-Small-24B and EuroLLM-9B also demonstrated improvements as the number of few-shot examples increased, though their absolute performance remained somewhat lower than Qwen2.5-14B across the scenarios tested.

The experimental results consistently show that adding task-specific examples boosts the model's detection performance. As the number of few-shot examples increased from 0-shot to 16-shot, most models improved their classification performance (see Figure 6), highlighting the crucial role that appropriate in-context learning can fill when applying general-purpose LLMs to specialized tasks.

Nevertheless, it was observed that models attained higher performance in classifying legitimate messages compared to smishing messages. This difference highlights an ongoing difficulty in using general-purpose LLMs to detect smishing. The models' weaker performance on smishing messages indicates they may have trouble picking up on the subtle hints, microtexts, or spelling that often characterize smishing messages. This finding opens the door to further exploration and refinement, possibly through focused fine-tuning and collection of more examples.

English versus Portuguese prompting As shown in Table 6, LLMs generally performed better when prompted in English, which is expected given their predominantly English training data. Nonetheless, some models, such as Qwen2.5 and Mistral, achieved results in Portuguese whose quality competes with those in English, reflecting the increasing multilingual capabilities of modern LLMs. English prompts also resulted in more stable and consistent improvements as the number of shots increased. In contrast, Portuguese prompts led to a decline in performance for models like Dolly, which exhibited notable fluctuations as the number of Portuguese shots increased. This contrast highlights Dolly's stronger alignment with English inputs.

6 Conclusion

In our study, we address an existing research gap in combating smishing attacks aimed at users of mobile money transfer platforms, specifically in a non-English context. To this end, we introduced a public, domain-specific, crowdsourced Portuguese language dataset designed explicitly for the task of detecting, and understanding smishing messages targeting mobile money users. Our exploratory data

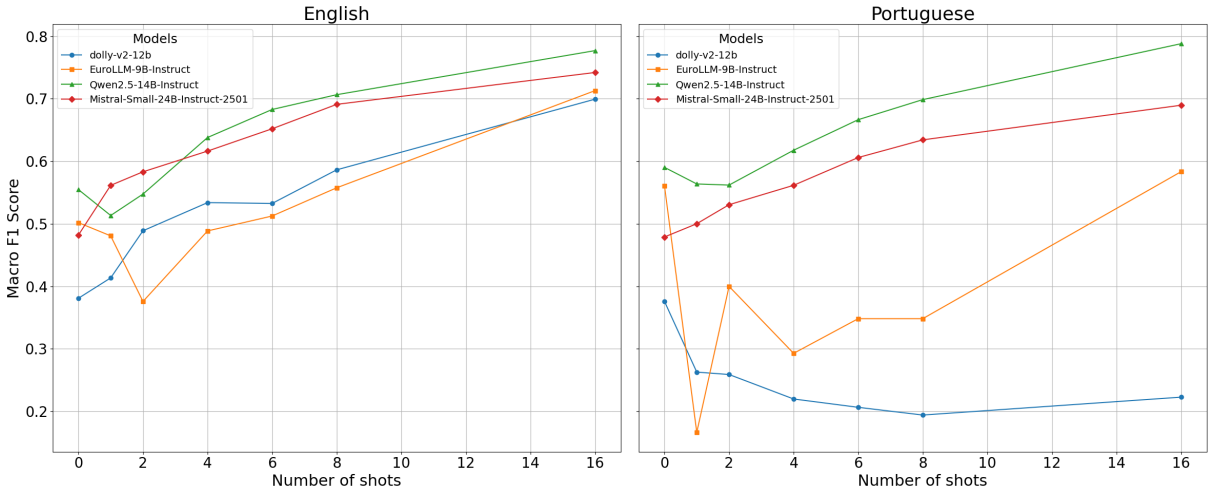

Figure 6: Macro F1 Scores for different models across few-Shot Settings

analysis revealed critical tactics and strategies employed by attackers, offering valuable insights that could facilitate the enhancement of user awareness campaigns and security tools.

Finally, our comprehensive experiments provided essential benchmarks evaluating how large language models perform through an in-context learning approach on this specific domain task. Our findings showed that models such as the multilingual Qwen2.5-14B demonstrated strong performance, particularly as more contextual examples were provided in the prompt scenarios.

Our research clearly underscores the potential of large language models to detect mobile money transfer fraud using careful task-oriented prompting strategies. However, the continued vulnerability of these platforms emphasizes a critical need for further training, fine-tuning domain-specific models, and improving general language AI capabilities to achieve greater sensitivity to linguistic nuances of text related to smudges.

Limitations

Despite the promising findings of this study, several critical limitations and constraints must be recognized:

Limited Computational Resources: The most significant limitation was the constrained computational capacity available through our hardware (4 NVIDIA A10 GPUs), which prevented us from experimenting with larger, state-of-the-art LLMs such as Llama-3.3-70B, Deepseek-R1-70B or Falcon. The inclusion of larger models may yield higher performances, but verifying this premise would require substantially larger computing resources than the ones at our disposal.

Lack of Temporal Dimension: Our dataset represents smishing messages collected within a specific time period and in the context of Mozambique. Thus, only static snapshot features of scams, which continually evolve, are captured. Further studies should capture longitudinal samples to track evolving fraud approaches and maintain effective detection.

Acknowledgments

This work was financially supported by UID/00027 - Artificial Intelligence and Computer Science Laboratory (LIACC), funded by Fundação para a Ciência e a Tecnologia (FCT), I.P./MCTES through national funds. Felermino Ali is supported by a PhD grant (with reference SFRH/BD/151435/2021), funded by FCT, as well as supported by the Base (UIDB/00022/2020) and Programmatic (UIDP/00022/2020) projects of the Centre for Linguistics of the University of Porto.

The authors thank everyone who contributed to the construction of the data set. Special thanks to Ekoko Clesh, Ednilson Sarmento, Clinton Uachave, Noémia Viegas, and all our colleagues and students at Lurio University, who have been actively conducting the data collection.

References

Tiago Almeida, José María Hidalgo, and Tiago Silva. 2013. Towards SMS spam filtering: Results under a new dataset.

Caner Balim and Efnan Sora Gunal. 2019. Automatic detection of smishing attacks by machine learning methods. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–3.

Tao Chen and Min-Yen Kan. 2012. Creating a live, public short message service corpus: the nus sms corpus. *Language Resources and Evaluation*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Dea Delvia Arifin, Shaufiah, and Moch. Arif Bijaksana. 2016. Enhancing spam detection on mobile phone short message service (sms) performance using fp-growth and naive bayes classifier. In *2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pages 80–84.

Abdallah Ghourabi. 2021. Sm-detector: A security model based on bert to detect smishing messages in mobile environments. *Concurrency and Computation: Practice and Experience*, 33(24):e6452.

Abdallah Ghourabi, Mahmood A. Mahmood, and Qusay M. Alzubi. 2020. A hybrid cnn-lstm model for sms spam detection in arabic and english messages. *Future Internet*, 12(9).

GSMA. 2024a. Mobile money fraud typologies and mitigation strategies. Technical report, GSMA.

GSMA. 2024b. The state of the industry report on mobile money 2024. Technical report, GSMA.

INTERPOL. 2020. Mobile money and organized crime in africa. Technical report, INTERPOL.

Ankit Kumar Jain and B. B. Gupta. 2019. Feature based approach for detection of smishing messages in the mobile environment. *J. Inf. Technol. Res.*, 12(2):17–35.

Ankit Kumar Jain and B.B. Gupta. 2018. Rule-based framework for detection of smishing messages in mobile environment. *Procedia Computer Science*, 125:617–623. The 6th International Conference on Smart Computing and Communications.

Ankit Kumar Jain, Sumit Kumar Yadav, and Neelam Choudhary. 2020. A novel approach to detect spam and smishing sms using machine learning techniques. *Int. J. E-Services Mob. Appl.*, 12(1):21–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Iddi S. Mambina, Jema D. Ndibwile, and Kisangiri F. Michael. 2022. Classifying swahili smishing attacks for mobile money users: A machine-learning approach. *IEEE Access*, 10:83061–83074.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Sandhya Mishra and Devpriya Soni. 2020. Smishing detector: A security model to detect smishing through sms content analysis and url behavior analysis. *Future Gener. Comput. Syst.*, 108:803–815.

Sandhya Mishra and Devpriya Soni. 2021. Dsmishsms-a system to detect smishing sms. *Neural computing & applications*, 108:1–18.

Pradeep Kumar Roy, Jyoti Prakash Singh, and Snehasish Banerjee. 2020. Deep learning to filter sms spam. *Future Generation Computer Systems*, 102:524–533.

Gunikhan Sonowal and K S Kuppusamy. 2018. SmiDCA: An Anti-Smishing Model with Machine Learning Approach. *The Computer Journal*, 61(8):1143–1157.

Daniel Timko and Muhammad Lutfor Rahman. 2024. Smishing dataset i: Phishing sms dataset from smishtank.com. In *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy*, CODASPY '24, page 289–294, New York, NY, USA. Association for Computing Machinery.

Kuldeep Yadav, Ponnurangam Kumaraguru, Atul Goyal, Ashish Gupta, and Vinayak Naik. 2011. Smsassassin: Crowdsourcing driven mobile-based system for sms spam filtering. In *Proceedings of the 12th Workshop on Mobile Computing Systems and Applications*, HotMobile '11, page 1–6, New York, NY, USA. Association for Computing Machinery.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

A Portuguese Prompt

Portuguese Prompt

A seguir estão exemplos de mensagens classificadas como Positivas (indicando intenção de smishing ou phishing) ou Negativas (indicando ausência de intenção de smishing ou phishing):

Input: "Bom dia, o valor melhor mandar para este nr 858798603 Mpesa nome Israel Robate Charimba, o meu atingiu limite."
Output: Positiva

Input: "Irmã peço pra me mandar mil mt."
Output: Negativa

...
Input: "Ok Aquele Valor Manda Para Este Nr D M-pesa 846861650 E Nome D Essinate Jofres"
Output:

Figure 7: Portuguese prompt template