Fortification learning to help detecting abuse in financial exchange description

Anuraj Bose
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Vellore, Tamil Nadu, India
anuraj.bose2022@vitstudent.ac.in

Dr. VIJESH JOE C
School of Computer Science and Engineering (SCOPE)
Vellore Institute of Technology
Vellore, India
vijesh.joe@vit.ac.in

***Abstract*— With the introduction of new payment methods and changes to the platform to allow longer texts in payment instructions, more people are already using communication systems. In some cases, this technology has also been used in targeted forms of domestic and family harassment. There are other challenges to overcome in detecting, responding to, and dealing with this type of technical support abuse.**

Keywords—Abuse, NLP, harsh language, and machine learning, Transaction details features, Simple text features.

I. INTRODUCTION (*HEADING 1*)

A. Abuse maked in banking enabled by technology

In the contemporary world, digital communication has emerged as increasingly vital. As of 2021, 4.55 billion people are active social media users, which comprises 57% of the weight of the global population[1]. The popularity and assortment of digital communication have offered us the capacity to contact someone 24 hours every day through a multitude of channels such as social media, text messaging, and email. Although this created greater convenience for many individuals, it has also brought significant challenges with regards to personal security and privacy, especially for those who are domestic violence victims/survivors[2].

The use of technological devices, such as mobile, the World Wide Web, or other forms of electronic communication, as a tool to enable individuals for carrying out abusive behaviours typically falls under the umbrella as technology-facilitated abuse.

such as coercive authority, intimidation, stalking, oversight, emotional and psychological assault, common bullying and unappreciated contact, sexual harassment, without the intention of triggering injury and suffering to the victim [3]. This expression can be expanded to encompass larger types of online harassment and cyberbullying, although it is primarily used in conjunction with gendered violence (domestic violence) [4]. Depression, worthlessness, exhaustion, self-harm, traumatisation, dread, isolation, emotional anguish, and other symptoms are possible consequences from technology-facilitated abuse. There have also been reports of economic punishments functional damages, and an intrusion on the recipient's personal freedom [4].

B. Banking abuse has been fostered by technology

Modern technologies for payment have quickened accelerated financial transactions while also allowing for greater descriptions of such transactions [5]. With the start of operation of the Bharat Interface for Money was developed by National Payments Corporation of India launched December 2016, where 1 a person or a business is able to send a real-time transfer to others and include up to 280 UTF-8 characters for reimbursement description and an extra 35 printed the American Society for Communication characters for payment reference.

Customers can additionally set up a simple being identified (UpiID®) for their accounts, such as a cellphone number or an email address, so they aren't burdened with recalling either Bank State Number (BSB)2 and account numbers every day. It resulted in a simple and quick method for people to communicate payments to one another in India.

As of the third month of 2022, 246 Indian financial institutions took advantage of these services, and consumers and business entities had registered over 35 million UpiIDs [6]. New technologies like UpiID simplified banking via increasing the convenience and total number of transactions, nevertheless they also gave those who commit crimes with an additional instrument for abuse.

The financial institution Commonwealth of Indian Banks, identified the usage of instantaneous interactions as a technique of improving operational effectiveness in early 2020.

That communication between people in particular, almost always via the technique of low value transactions. Over a three-month period, thus discovered that over more than 17,000 clients had received many states low-value deposits with possibly offensive remarks in the transaction description. We selected that the communications' aim ranged from "jokes" containing profanity to severe threats or references to domestic misconduct while participating family violence [5].

Using transaction descriptions for inappropriate communication or misuse rather than to transfer payments has been confirmed by banking institutions all throughout the world. Depending to the Australian Transaction Reports and Analysis Centre (AUSTRAC) Fintel Alliance investigate [5] it is not unique to the Australian banking industry. Several Brazilian news outlets, for example, discovered the arrest of an individuals who had been harassing a young woman via bank transfers after his phone number was removed from the database [7]. The players can see that any financial instrument that has a free text area to be filled up by the sender and read by the receiver is vulnerable to being utilised to facilitate illicit communication.

C. Role of the Bank

It is not unique to the Australian banking industry, according to the Australian Transaction Reports and

Analysis Centre (AUSTRAC) Fintel Alliance study [5]. Several Brazilian media organizations, for example, report the arrest of a guy who was harassing a young woman through bank transfers after his phone number was banned [7]. We can see that any payment that has a free text form that the sender needs to complete up and the recipient must view can be a vehicle for illicit communication.

Although these approaches have mitigated the use of profanity in online banking transactions, they are not deterring abusers who want to inflict injury or distress to victims since they have simply learned how to overcome these first measures. For example, the phrase unblocks, which has been associated with these abusive payments, has been reported to be transformed to un-block, u.n.b.l.o.c.k., and other modifications in order to circumvent it. As a result of these developments, we decided to protect our consumers by developing a monitoring system that may operate in the background, comprehending the circumstances of significant improper use that may require additional investigation.

The definition of exploitation and the emergence of mechanisms and processes that might benefit victims and deter individuals comprise just two of the many components that go into developing a system that is capable of detecting abuse. We must first agree these incidents in order to decisively stop the abusers or communicate with the influences clients and offer help. The substantial number of transactions sent each day, the complexity of comprehending the context of each transaction passed around, and the delicacy of the language and conduct employed all contribute to the having trouble of this task alone. We capitalized a multi-step marketing to improve resolving the issue that was occurring. The model is used to score all of the transactions first. A specifically group of customer vulnerability specialists examines the cases with the highest accomplishes the feat scrutinizes them manually, and gets throughout touch with any identified abuse victims. The team will then consider their most appropriate course of action, which may include instructing the victim and writing letters to the offender providing them that their actions is not satisfactory. In specific situations, welfare checks may be worthwhile to make sure they are safe and to get authorization from them to take further action. The team is permitted to manually measure and process an authorized number of cases on a monthly basis due to the capacity and complexity of the cases and solutions given. In order to ensure that we determine all of the real positive cases, it is indispensable that we limit the overall number of false positive cases. The current technique as much as possible the one that we get could have been improved by efforts throughout the more general research community, fails to incorporate comparison with alternative models irrespective of the distinctive characteristics of the subject being investigated. But the new work creates an authoritative basis for for comparison. We also believe that using these methods could benefit other kinds of banks that now use basic filters and keyword recognition that are insignificant to go anywhere.

II. Problem Statement

The researchers who participated in the present investigation offer another approach that recognizes high-risk misuse cases in banking payment systems by combining information from multiple states deep learning models. Technology-assisted abuse is not an unresolved problem, and research has been done to explore it (see Section III above for additional information), but this particular misuse employing financial transactions was just recently discovered and presents a new set of difficulties. The sensitive nature of the subject is one of the main obstacles. It should be approached with the greatest of reservations since both taking action and doing nothing have the potential to be harmful. Unlike message boards messages that may be readily removed or banned, bank transactions are unable to defend. The transactions' "life-span" has become substantially longer. They might be seen by someone besides the victim. They might be communicated in written form. They may be cited as encouragement in applications for loans alongside other services, which could lead to re-traumatization. These circumstances can be challenging for someone to manage since the victim may be significantly fewer tolerant of the abuser's behaviors.

In regards to the steps implemented, a financial institution often contacts customers and demanded that they cease. If the improper conduct persists, a client may be barred from banking. Un-banking, in comparison to social media bans, is the choice of an insurance company to cease its connection with some particular individual and may have major repercussions on the way individuals live. Transaction descriptions frequently contain constrained context and are vulnerable to interpretation, which only serves further confuse matters further. As a result, each case must be inquired about and approached separately by a specifically designed team, a process that demands a lot of effort and accomplishments in the concentrating on of the occasions.

III. LITARATURE REVIEW

A thorough examination of all of the technologies used by abusers when committing technology-facilitated abuse anchored to domestic violence has been provided in a study by [11]. The paper addresses numerous kinds of abuse related to coercive control, financial abuse, smart homes, and stalking, as well as how abusers exploit use of these. The study additionally presents a framework for inclusive safety in technology system design, but it presents no recommendations at all for ways to spot when a system is being abused. Although the paper emphasizes how financial abuse might take place in banking systems, it neglects to address the issue of abusers leveraging transaction descriptions in order to communicate abusive messages and engage in controlling and stalking behaviors.

Finding comments that are inappropriate in bank transaction descriptions is a brand-new concern. While

studies concerning similar problems, for example discovering foul language, toxic levels in text, bullying, and hate speech, has been currently underway for the past 20 years, it has primarily concentrated on social network moderation. As an example, machine learning techniques have been used to detect this kind of content on Twitter [12] and Facebook [13].

Similar to the issues associated with abusive communications, social network data is predominantly informal, unstructured, and sometimes misspelt. As a result, documents like [14] have used algorithms for natural language processing that recognise both lexical and morphological aspects of sentences. [14] implemented style, structure, and posting pattern considerations in addition to the text's contents to increase their recognition of unconstitutional posts. Among other characteristics, [15] used emoticons, uppercase, the amount of followers, and relief.

[13] introduces an unusual approach that is referred to Entailment as Few Shot Learner (EFL). This approach involves turning class labels through a natural language phrase that is used to describe the label and determining if the label contains the description in order to enhance language models as few-shot learners. The EFL approach might be broadened to multilingual few-shot learning and can additionally make use of comes to like unsupervised contrastive data augmentation. [16] demonstrates an uncommon shallow neural network to classify whether comments are undesirable or instances of assaulting in a cyberbullying the surrounding environment using GloVe embeddings on public Wikipedia datasets.

To distinguish amongst targeting and untargeted unfriendly words, [12] made use of machine learning. For the purpose of to do this, a three-level annotation schema that corresponds to the three subtasks has been established. The first Subtask A was merely concerned with understanding the dataset's language that was antagonistic or not inappropriate. likewise, Subtask B considered the data as hate speech or general undesirable language, or as focused on or untargeted, and Subtask C organised the data in accordance with whether or not the hate speech was directed at a particular person or a group.

Other technologies for discerning abuse have utilised the use of pre-trained language model-based systems like RoBERTa and BERT [17], which have achieved new state-of-the-art results on a broad spectrum of tasks [18]. The Twitter Hate speech and Wikipedia databases have been utilised in [19] to identify abusive language by utilising a BERT model that has been fine-tuned using binary cross-entropy loss. When compared to other the embeddings including fastText, TextCNN, and TextCNN + Character n-grams, BERT embedded models performed well superior. Pre-trained models have challenges on domain-specific language tasks due to the fact that they were trained on common datasets, which is a problem. The standard answer to this is to retrain already-mastered models by employing domain-specific datasets, as presented in [20]. In circumstance like the detection using abusive language, because there is inadequate effort data necessary for creating from scratch a BERT-like model. Their model, called "TweetBERT," fared more favourably than being there BERT-based models when investigating Twitter material after having been retrained on a Twitter-based corpus.

A comparable scheme for enhancing BERT-based models has been provided in Paper [21] in order to recognise instances of cyberbullying and abusive speech using Twitter data sourced from Australia. This has been accomplished by adding extra functionality as unique tokens to BERT. Emoji pathways were inside the features, in conjunction with metadata consisting of user information (e.g., gender, year of birth, number of posts), information about their network (e.g., number of followers and friends), and information about their influence (followers/friends ratio). According to the results, BERT with the additional characters (BERT + emoji + network + power) generated the highest level of precision.

We discovered that previous research focuses on more locating instances of abusive relationships than its focus on the abusive relationship. Such transactions descriptions require to be more frequently supplied and at least somewhat uniform. Similar to this, additional papers have produced methods to detect potentially abusive users rather than occasional instances of abusive language. As an illustration, [22] employed graph machine learning in order to identify nasty those involved. Hateful accounts have been characterised using, among other things, factors including formation date, user activity, network centrality, passion, and a linguistic examination. The technique entailed deciding on users from a neighbourhood and organising them as hateful or non-hateful using an application based on DeGroot's learning model [22]. The word important trends, such as greater activity and a greater likelihood of using certain terminologies, and have been determined to be correlated with "hateful users."

There has been no study in the direction of understanding abuse in transaction descriptions in the context of financial services, despite having been discovered that there is a lot of work that revolves around online social networks. In the present investigation, we make use of a number of machine learning approaches to not only spot abusive language in the transaction descriptions but also to pinpoint the customer connections involved in such transactions.

IV. DATA

Thus, this paper presents the specific features of the dataset we make use of in the second section, and we also describe the methods utilised for data pre-treatment before use. Transaction information gathered from the Commonwealth Bank has been employed in the following paragraphs. This paper discovered the transaction descriptions, their associated dollar amounts, the date of the transaction, and the person who sent and received the account numbers. data extracted from the bank's database. This paper produced information concerning transactions from both the non-NPP procedures and the newly introduced payment platform (NPP). In compliance with Section V, this data was utilised to create features for the machine learning training that this paperer aggregated by relationships. Understand that a relationship in the present instance refers to a sender and recipient pair of a transaction; for instance, in the event that originator a dispatches a transaction to the person receiving it, this paper have an existing connection a, b; nonetheless, if recipient b communicates a transaction to a, this paper have a new and distinct relationship b, a. The preceding time-window this paper utilised determines how

many transactions this paper used for each connection. This paper set the time span for this investigation to be one month in length. 1,039 relationships in our dataset have been identified as either (1) substantially destructive or (0) non-abusive. The standard measure of high-risk abuse (see specifications 2.1) had been used by multiple domestic violence specialists when classifying 283 of those unusual arrangements as "highly abusive."

They recorded 87% in agreement. haphazardly deciding on non-abusive attachments and a random selection of circumstances at which expenditures had "conversational" characterizations that this paper compliant with the abuse perceptions but this paper most especially different from normal interactions made it possible an experiment of negative sampling. In numerous instances, potential customers sent one another song lyrics or involved themselves in a completely ordinary conversation. In order to stay away from a model trained by machine learning from launching up just lengthy messages without considering highly hazardous misuse, this needed to be done. The data in this training set ranges from July 2021 to the first month of 2022. Thus it's conducted our investigations using this dataset, and thus implemented k-fold cross validation to validate the solution thus had suggested. There are no overlapping combination pairs betthis paperen folds, it is very important for one to understand this.

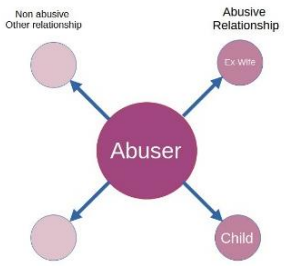

Figure 1:Some abusers frightening many different kinds of victims. In the present case, the communications in between the abuser and his kid have been considered on their own in our dataset from the abuser's relationships with his ex-wife. **It is important to into consideration that the transaction descriptions in this graphic are fictional and were created to show how offensive these remarks are.**

On top of that, thus it is discovered transaction data for a month from a previously collected out-of-sample data sets. For this, thus worked with information from the entirety of the month of February 2022. This functions as a glimpse of a model scoring use-case the interior of an existing business process. Less than 0.0005% of scenarios in a particular month are abusive, establishing a significantly disproportionate problem. The out-of-sample test set was assembled by categorising the top fifty highest obtained relationships of the month that corresponded for each of the candidate models, 35 of which turned out to be extremely abusive, as manually rating all of these monthly transactions is impracticable.

V. METHODOLOGY

Additional information concerning the plan thus have is made available in this section. Choosing whether to look identify abuse at the transactional, customer, or relationship level was the first challenge. Insufficient textual data prevents the transaction level from capturing context. Think of a text message that simply states, "I love you." It's difficult to tell whether this is a case of harassment or just a routine communication between a pair without further transactions to study the dynamics. However, the situation becomes much evident if this kind of explanation is delivered every five minutes and the other side wants to stop. However, the offensive information may be diluted if this paper takes into account gathering all transactions at the client level. As an example, despite having an extensive collection of dependable recipients, determined customers consistently annoy just a handful of them wherever they go. Because of this, by leveraging descriptions achieved from purchases between each sender-recipient pair, this can identify misuse at the relationship level of responsibility. An abuser with a variety of victims can be observed in Figure 1, and in this particular situation, this will recognise the two separate relationships in which they are at high risk.

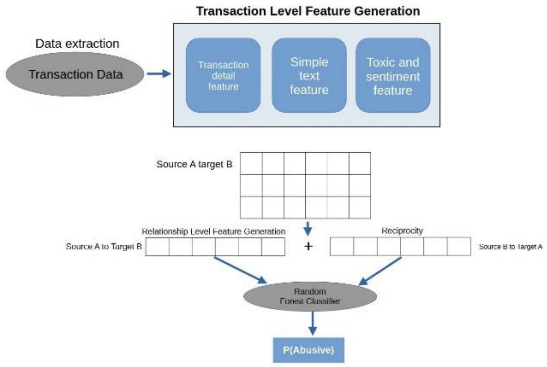

Figure 2 brief explanation of the architecture of the system's components. The unstructured information has been turned into three sets of traits, ranging which are subsequently incorporated in accordance with their relationship intensities. The final model inc.

Thus it has immediately go through the overarching approach constructed for distinguishing fraud in expenditure descriptions (AITD). It should be accentuated that our primary objective is to find the most egregious situations. The general description of our system is illustrated in Figure 2 and encompasses the following steps:

1. Generating appropriate characteristics from every individual transaction description is commonly referred to as "transaction-level feature generation" (Section V-A).
2. In order to uncover abusive customers while excluding just the particular abusive transactions (Section V-B), relationship-level feature manufacturing involves aggregating the aforementioned characteristics on every interaction (sender-recipient pair).
3. Considering reciprocity information: constructing features dependent on the almost certain victim's prospective responses (Section V-C)
4. Constructing a computational model to foretell the labels: Relationships are were classified as either extremely harmful or non-abusive by employing the random forest model.

Feelings, the album Toxicity and Temperament characteristics, which are features based on three language models that were previously already trained and have the capacity of delivering helpful details for the abuse the identification predicated on the language used in the communication summaries.

Using an independent the variation of the BERT-based, previously trained linguistic model Detoxify, seven toxicity the characteristics have been determined for every interaction [23]. The Detoxify the program's desired outcome version differentiates poisonous substances and reduces unintentionally who you are bias. It determines the text's harmful effects in seven different categories, which includes adverse effects, severe toxicity, obscene speech threat, insult, identity domestic violence, and sexual degree of disclosure. The results of this assessment were subsequently took advantage of for establishing the seven toxicity the characteristics for the identified AITD model. This implemented DistilBERT as models prepared on four data sources, which includes the daily dialog, empathy-stimulus, isear, and hugging face emotion datasets [24], for determining emotion defining features. This pre-trained model determines if the uploaded text has resulted in the emotions regarding love, fear, surprise, or is neutral, cheerful, sorrowful, or dissatisfied for every single interaction advertisement, it is assumed the review scores for each emotion component. In this our purpose of discovering the sentiment of a transaction description, it is implemented VADER, which stands for Value-Aware Dictionary for Sentiment Reasoning). It communicates both the amount of force (the level of intensity and polarity (positive/negative) of the feeling in question. Actual emotions analysis of the VADER is based on a dictionary that collaborators vocabulary words with sentiment scores, which are a measure of the degree of intensity of an emotions. The degree of significance regarding every single word with a text can be introduced to come up with the sentiment score. The proposed Sentiment Intensity The analyser was for the VADER algorithm takes a text feedback and results in a dictionary of numbers in four categories, which might involve favourable, unfavourable, the substance, and neutrality [25].

B. Production of relationship-level features

The starting point, as previously responded, is the creation of capabilities at the transaction surface. This transaction-level vector of attributes is subsequently merged to provide relationship-level features. In our model, three different types of attributes can be observed:

- Transaction details characteristics (TRX) are excluding to the individual details of the transaction between sender and receiver and encompass the dollar value of the amount performed, purchases date, the total number of agreements per day, maximum number of transactions per day, and the interval between the highest and lowest possible number of transactions.
- Simple textual characteristics (ST) referred to the simple information the fact that can be obtained through the transaction itself overview, which include the length of the expenditure description, upper/lower/mixed case flags, the total amount of words, length of the word that is longest in the transaction description, the extent to which the content of the message possesses special characters or numbers, depleted an explanation emblem, different periods grammar, punctuation and number-related flags.

Table 1 : TYPE OF AGGREGATION FOR FEATURES. Every action taken within a relationship are susceptible to feature aggregation.

Aggregation	Features
Maximum	Every sentimental element
Minimum, Maximum, Median	Length of transactions, number of words, the longest word length, proportion of word breaks to the message description length
Sum	All the toxicity features
Mean	All the emotion features, transaction amount, number of lower case words, number of upper case words, number of mixed case words, number of punctuation found

Every single transaction was calculated to include the features mentioned above. This has to bring together the information gathered from every communication amongst a sender and an intended recipient since the primary objective of our estimation task is links. a range of the properties employed, as reported in Table 1, the collection is carried out in a slightly different manner.

Further, this created advantage of the characteristics that have been gained from each engagement in an alliance. These are the aggregate amount of operations sent throughout the development of a couple's relationship, the greatest amount of purchases ever communicated in one single day, and the total a number of distinct days on which these actions have occurred place.

Table 2 Emotion, toxicity, and sentiment (ETS), simple text (ST), and transaction (TRX) feature combinations: experimental results.

Features			***Performance***				
ETS	***ST***	***TRX***	***Prec***	***Rec***	***F1***	***AUCPR***	***ROC AUC***
■			0.618	0.740	0.670	0.526	0.766
	■		0.615	0.686	0.645	0.505	0.748
		■	0.575	0.666	0.614	0.474	0.731
■	■		0.633	0.728	0.671	0.531	0.775
	■	■	0.657	0.721	0.683	0.547	0.790
■		■	6.638	0.709	0.669	0.532	0.780
■	■	■	0.659	0.730	0.690	0.554	0.795

Table 3 : EXPERIMENTAL Performance Included RECIPROCITY PROVIDED TO THE most effective

MODEL INCLUDING EMOTION, TOXICITY, AND SENTIMENT (ETS), SIMPLE TEXT (ST), AND Payment (TRX) Characteristics.

Prec	Rec	F1	AUC-PR	ROC AUC
0.678	0.738	0.703	0.570	0.800

C Reciprocity

It is also add the same attributes that are allocated by an interaction's responses received. In another way, characteristics on the connection between "a, b" are taken into consideration when combined with features on the "b, a" reciprocal relationship. This concurs with our contention that reciprocity might have been positive since a recipient will constantly refrain coming from responding to a determined sender.

VI. Result and Discussion

Firstly, it is undertook a research study to figure out which traits would be most appropriate for differentiation between occasions involving severe violence and non-abusive occasions. Each of the following set of characteristic a combination have been employed for assessing the models: information about the transaction (TRX), simple text (ST), and toxicity and sentiment characteristics (ETS) Table 2displays the outcomes of continuous 5-fold cross-validation, and metrics for measurement which includes precision, recall, F1, AUC, and AUC-PR. As a whole, the random forest approach with amalgamated plain text, trans-action particulars, feelings, as toxicity, and emotion parameters (ETS + ST + TRX) experienced the best performances. This will conducted experiments with adding mutually beneficial features after settling on the best sets of the characteristics and further changes were (seen in Table 3).

Then, considering the out-of-sample test set referred to in Section IV, this considered our findings. For demonstration purposes the power of our model, thus employed the top system retrieved from the previous investigation (ETS + ST + TRX + reciprocity). In order to efficiently and efficiently designate the top episodes for a manual evaluation, thus necessary to be bound that the technique was reliable and that it had not produced any false positives for those instances with the highest scores.

Assuming no overlap in this month's data and the ones from our training dataset, Table 3 shows the ROC curve of the models on the out-of-sample test set of the transaction descriptions accumulated all throughout one month.

The highest 50 sender-recipient pair circumstances are then completely associated and then utilised for developing the ROC curve. The additional cases, and these contain several hundred details about the transaction and incorporate a lot of human place of employment, had not been manually checked.

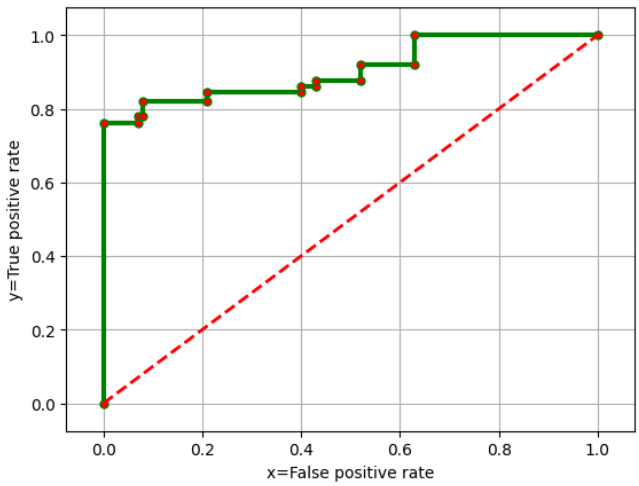

Figure 3 : The ROC curve (green line) for the group of fifty situations the fact that have the highest corresponding probabilities of the out-of-sample set being present, as calculated using our best model (ETS+ST+TRX) + reciprocity.

The receiver operating spectrum (ROC) curve displays the trade-off involving specificity (which is one minus false positive rate) and responsiveness (true positive rate or recall). In Figure 3, a black dotted line denotes a rough figure. It ought to been noted that classifications with curved that are more nearby to the top-left corner performed better for instances with the highest scores. Fig. 3 highlights thoroughly abundantly clear that the best system reliably forecasts the very hazardous scenarios with its initial true negative observed in the 26th case.

Conclution

In this piece of writing, this has addressed a fresh problem involving intimidation and assault in the finance industry domain names. Though sharing a few similarities with other social networks on the web, this particular problem presents additional challenges and requirements for special attention. It is talked about a specific an example of improper utilisation of transaction characterizations in Australia's largest bank and gave solutions. This analysed models with various sets of features and evaluated how well them carried out in a real-world situation. Thus established that the model with the greatest performance is an automated model trained on a variety of features, stretching in complexity from simple bending transaction and text the characteristics to toxicities and responses variables that are computed utilising advances in the field of NLP. The bank is now employing the model as it is in its final form. while the examples supplied can be verified by the client being vulnerable specialists, we repeatedly retrain the model in order to boost its robustness.

In order to preserve and secure the people thus serve, this regularly upgrade our system. This have been working on several types of variations and anticipate unveiling these with your our forthcoming projects. Better international language policy coverage, by utilising conversations that occur over lengthy periods for spotting long-term abuse, and incorporating BERT embeddings for the detection of high-level attributes when the designated as training set is sufficiently big are some instances of potential updates.

REFERENCES

[1] Social, We Are. "Social Media Users Pass the 4.5 Billion Mark." Retrieved September 23 (2021): 2021, https://wearesocial.com/au/blog/2021/10/social-media-users-pass-the-4-5-billion-mark/

[2] Dragiewicz, Molly, Bridget Harris, Delanie Woodlock, Michael Salter, Helen Easton, Angela Lynch, Helen Campbell, Jhan Leach, and Lulu Milne. "Domestic Violence and Communication Technology: Survivor Experiences of Intrusion, Surveillance, and Identity Crime." (2019).

[3] Fiolet, Renee, Cynthia Brown, Molly Wellington, Karen Bentley, and Kelsey Hegarty. "Exploring the Impact of Technology-Facilitated Abuse and Its Relationship with Domestic Violence: A Qualitative Study on Experts' Perceptions." *Global qualitative nursing research* 8 (2021): 23333936211028176.

[4] Flynn, ASHER, Anastasia Powell, and Sophie Hindes. *Technology-Facilitated Abuse: A Survey of Support Services Stakeholders*. ANROWS Research Report, 2021.

[5] "Preventing Misuse and Criminal Communication through Payment Text Fields." Accessed 9 Dec 2021, 2021. https://www.austrac.gov.au/business/how-comply-guidance-and-resources/guidance-resources/payment-text-fields.

[6] Leontjeva, Anna, Genevieve Richards, Kaavya Sriskandaraja, Jessica Perchman, and Luiz Pizzato. "Detection of Abuse in Financial Transaction Descriptions Using Machine Learning." *arXiv preprint arXiv:2303.08016* (2023).

[7] UOL. "Homem ´E Preso Suspeito De Fazer S´Erie De Pix De Centavos Com Ameac¸as `a Ex." Last modified 20-06-2022, 2022. https://noticias.uol.com.br/cotidiano/ultimas-noticias/2022/06/20/homem-e-preso-suspeito-de-fazer-pix-com-ameacas-a-ex.htm.

[8] Kane, Annie. "Major Banks Reveal Major Issue of Abusive Transactions." The Adviser. 2021. https://www.theadviser.com.au/breaking-news/41494-major-banks-reveal-major-issue-of-abusive-transaction.

[9] Loop, H. in the. "What Is a Human in the Loop?". 2020. https://humansintheloop.org/what-is-a-human-in-the-loop/

[10] Resources, Australian Government Department of Industry and. 2022. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

[11] PenzeyMoog, Eva, and Danielle C Slakoff. "As Technology Evolves, So Does Domestic Violence: Modern-Day Tech Abuse and Possible Solutions." In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*: Emerald Publishing Limited, 2021.

[12] Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. "Semeval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (Offenseval)." *arXiv preprint arXiv:1903.08983* (2019).

[13] Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. *Detecting Offensive Language in Social Media to Protect Adolescent Online Safety*. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing: IEEE, 2012.

[14] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012: IEEE, pp. 71-80.

[15] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Measuring# GamerGate: A tale of hate, sexism, and bullying," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 1285-1290.

[16] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[19] S. B. Bodapati, S. Gella, K. Bhattacharjee, and Y. Al-Onaizan, "Neural word decomposition models for abusive language detection," *arXiv preprint arXiv:1910.01043*, 2019.

[20] N. Azzouza, K. Akli-Astouati, and R. Ibrahim, "Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations," in *Emerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing 4*, 2020: Springer, pp. 428-437.

[21] A. Leontjeva, G. Richards, K. Sriskandaraja, J. Perchman, and L. Pizzato, "Detection of Abuse in Financial Transaction Descriptions Using Machine Learning," *arXiv preprint arXiv:2303.08016*, 2023.

[22] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. Almeida, and W. Meira Jr, "" Like Sheep Among Wolves": Characterizing Hateful Users on Twitter," *arXiv preprint arXiv:1801.00317*, 2017.

[23] L. Hanu, Ed. *Detoxify*. 2020. [Online]. Available: https://github.com/unitaryai/detoxify.

[24] B. Savani, "Distilbert-base-uncased-emotion model," 2020. [Online]. Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

[25] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014, vol. 8, no. 1, pp. 216-225.