Development of a Risk Warning System Using Limited Data

Morris John Montemayor, Ron Emil G. Castro

Azcalun Inc.

Author Note

Data collection and preliminary analysis were performed by Azcalun Inc., with a supporting grant from the Department of Science and Technology–Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD).

Azcalun Inc. has developed platform features based on this study. The machine learning model remains active and is continuing to learn based on current chat activity.

Table of Contents

Abstract

This research explores whether it is possible to produce a machine-learning model to predict risky conversations based on the chat history of fewer than 20,000 users.

Keywords: risk warning, scam detection, risk prediction, text chat

Development of a Risk Warning System Using Limited Data

Online Social Networks (OSNs) derive value from connecting users to one another. However, the value that is created is put at risk when the networks are used by unscrupulous users to perpetuate scams against other users.

Aldwairi and Tawalbeh (2020) describe how malicious activities over OSNs can be detected using the social graphs of connections between users to determine trustworthiness based on relationships (Graph-based Analysis) or by relying on feedback from users who report malicious activity (Manual Verification). They also describe using Machine Learning to develop artificial intelligence machines that can detect malicious activities.

The research team attempted to use available data within the kazam™ platform to train Machine Learning Models to detect potentially fraudulent behavior.

Training Cycle 1

Message content from the kazam™ platform was extracted into secure environments to create datasets for machine learning purposes. The data was associated with manual user reports of abusive behavior to determine which messages were sent by users who were banned after a manual report of abusive behavior vs. those that were sent by users who have not been banned. The segmented data was then pre-processed and further segmented into separate datasets for training and testing purposes.

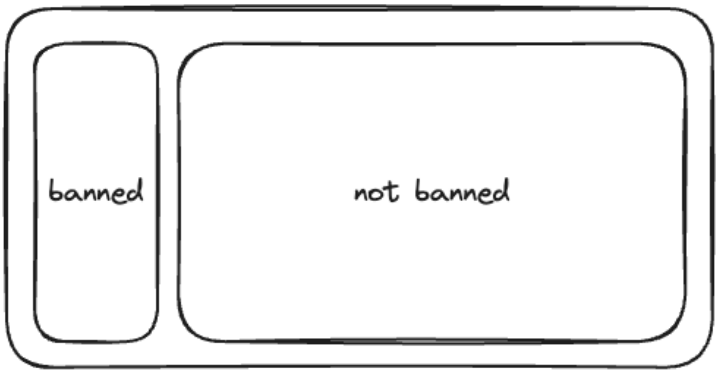
Pre-processing

The raw message extracts were first put through an initial pass to exclude data points that could skew the learning process. Aside from duplicate messages (i.e., messages sent by a user into multiple chats), the team also excluded a templated auto-generated message sent by the platform to start each chat session as well as any chats between users and the kazam™ administrator account. The latter was excluded as the interactions between users and the kazam™ admin account were very different in nature and content and would therefore not be comparable to the user-to-user interaction that the machine learning models are intended to learn.

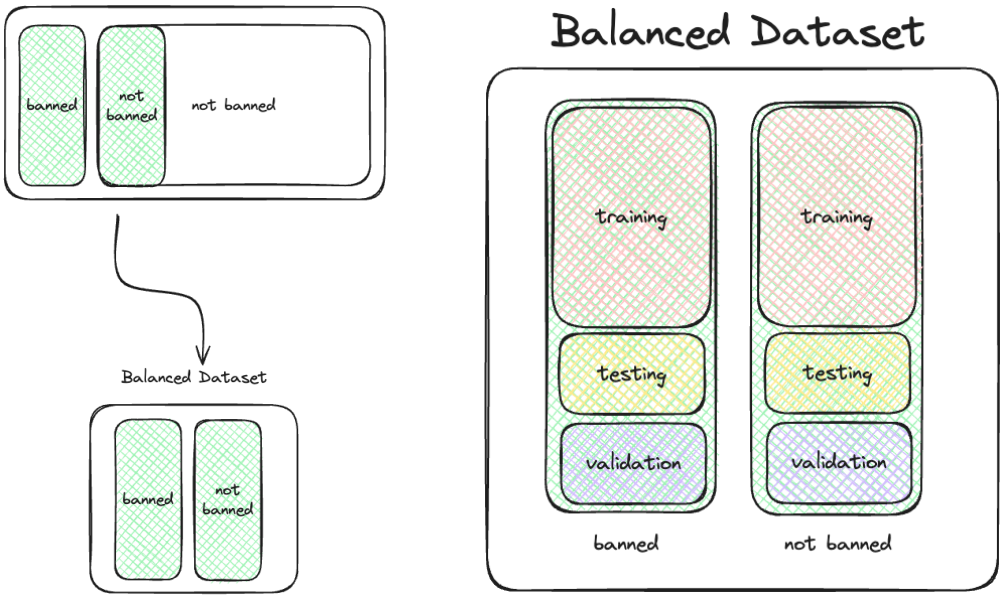
Using a Calamancy model-based transformer and regex pattern matching, privately identifiable information (PII) in message content was replaced by static tags to avoid exposure of the information (e.g., “<NAME>”, “<NUMBER>”, “<LOCATION>”, “<EMAIL>”, and “<URL>”). The pre-filtered and sanitized data was then segmented based on whether the users that sent the messages were banned or not banned.

Pre-processing yielded a total of 4,185 messages from 49 banned users and 39,694 coming from 2,711 users who were not banned.

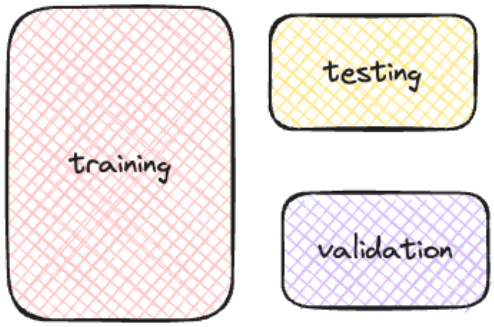
Dataset Preparation



Pre-processing yielded a dataset where messages associated with users who were not banned for abusive behavior are over-represented by an order of magnitude. To deal with this, random sampling was done on the data associated with users who were not banned to produce a sample of equivalent size to the banned users dataset. From the balanced dataset, another random sampling was made to performed to split the balanced dataset into three datasets: training, testing, and validation.



This sampling yielded a total Training dataset of 3,366 messages, a testing dataset of 1,122 messages, and a validation dataset of 1,122 messages. The training dataset was used to train the model. The testing dataset was used to fine-tune the hyper-parameters of the model. The validation dataset was used to test the fine-tuned model.



Model Training and Validation

The team created a FastText model to predict whether a user was likely to end up being a banned user based on their messages. The model was designed to look at each message individually, with the goal of reviewing each message in a stream and triggering a warning message for any detected positives. Calamancy (calamancy-tl-lg-0.1.0) was used for feature extraction and XGBoost was used for classification. Optuna was used to optimize the hyperparameter of the XGBoost model across 150 trials. The training and optimization phase of the model yielded an accuracy score of 0.8673170731707317.

Running the validation dataset against the model yielded the following results:

Metric	Value
Accuracy	0.89
Precision	0.83
Recall	0.83
F1 Score	0.83
AUC	0.87

- Accuracy: 0.89 *(ratio of correctly predicted observations to the total observations)*
- Precision: 0.83 *(the ratio of correctly predicted positive observations to the total predicted positives)*
- Recall: 0.83 *(also known as sensitivity, is the ratio of correctly predicted positive observations to all observations in actual class)*
- F1 Score: 0.83 *(the harmonic mean of precision and recall that shows a balance between the two metric)*
- AUC: 0.87 *(Area Under the receiver operating characteristic Curve assess the performance of a binary classifier in distinguishing positive and negative classes)*

Training Cycle 2

As the scores from Cycle 1 were lower than expected, the team revised our approach to consider the entire conversation instead of each individual message. A step that concatenated all messages sent by a user in a chat conversation was added to pre-processing. The change in approach allowed the model to achieve higher scores across all metrics.

Metric	Value
Accuracy	0.93
Precision	0.92
Recall	0.87
F1 Score	0.90
AUC	0.92

The team deemed this version of the model acceptable to release as a feature in the kazam™ platform.

Results

After months in production, the model performed as follows:

- Out of a total of 1,913 chats from 108 users who were flagged by the model as having risky conversations:
    - 156 chats were flagged by the model, out of which only 2 were manually flagged by users as abusive.
    - 10 chats from 10 users were manually flagged by users as abusive without those chats having been flagged by the model as risky
    - Out of these 108 users, only 1 ended up being Banned and 5 being Suspended.
- Out of a total of 2,927 chats from 37 users who were manually reported as abusive:
    - 31 chats were flagged by the model, out of which only 2 were manually reported by a user as abusive

- 38 chats were manually reported as abusive, without having been flagged by the model as risky
- Out of these 37 users, 4 ended up being Banned and 12 ended up being Suspended. Out of these 16, only 6 had at least 1 chat that was flagged by the model.

On one hand, only 1.9% of the users flagged as having risky behavior were validated as such by a user manual report[i]. On the other hand, 37.5% of the manually reported abusive users were validated by the model as having risky behavior. This performance implies some limited success for the model, albeit not enough to declare that it is able to predict scams with a high level of confidence.

References

Aldwairi, M., & Tawalbeh, L. I. (2020). Security techniques for intelligent spam sensing and anomaly detection in online social platforms. *International Journal of Electrical and Computer Engineering*. https://doi.org/10.11591/ijece.v10i1.pp275-287

Castro, R. E. G., & Azcalun Inc. (2023). *Theoretical Framework: Improving the safety of users on Online Social Networks*.

Foody GM (2023) Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *PLoS ONE 18(10): e0291908. https://doi.org/10.1371/journal.pone.0291908*

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics, 61(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x*

---

[i] There is currently no incentive for users to report abusive behavior and no mechanism for them to confirm that they deem the chat that was flagged as risky as being not risky. The platform may need to be enhanced to provide additional instrumentation to allow users to vote for flagged chats as being non-risky to confirm false-positives.