TEAM 3

PROJECT PROPOSAL

10/12/2024

Prepared For :
Necdet Gurkan

UMSL
Campus office 234

PROJECT TEAM 3

Jeremiah Fowler

Tim Burke

Navya Rani Kanaka

MJ West

Jin Bai

OVERVIEW

The purpose of this project is to enhance our customer's experience with a superior client by reducing spam and phishing text/SMS messages. We plan to implement a robust model with the objective of accurately detecting and blocking unwanted messages, either on backbone nodes, or directly on end user devices. The scope of this project is focused on SMS but could be expanded to include other more feature rich text services or serve as a framework for an audio detecting model. The expected outcome is to reduce unnecessary messages to our client base.

The value of this project is in the reduction of unwanted messages for our customer, positioning this feature as a unique differentiator in our product offerings. This feature will positively impact our cumulative margin, as while texts are allowed on an unlimited basis, adding this service as an add on feature can reduce over spam and phishing messages clients receive, which will lower customer support costs and improve service reliability.

BACKGROUND & IMPORTANCE

This project aims to enhance customer experience through the reduction of spam and phishing messages, addressing a major frustration for users while safeguarding them from harmful scams. By providing a secure, trustworthy messaging experience, we will gain a competitive edge, boosting customer satisfaction, loyalty, and brand reputation.

The rise in phishing and scam messages burdens our customers financially, emotionally, and wastes their time. This in turn drives up expenses for sectors like insurance and financial services that manage these risks. The development of a reliable "Spam or Ham" detection model is essential to reduce fraud, protect customers, and reinforce our commitment to user safety, and ensure peace of mind for customers engaging with our messaging service.

PROBLEM STATEMENT

The goal of this project is to develop a machine learning model that accurately and precisely categorizes incoming text messages as either spam or ham (legitimate messages). This solution aims to reduce the burden placed on individual users to discern whether a text message is legitimate or not. By implementing our robust model, we are able to support both individuals and organizations in filtering our potential scams efficiently and reliably.

TECHNICAL SPECIFICATIONS

The goal of this project is to develop a machine learning model that accurately and precisely categorizes incoming text messages as either spam or ham (legitimate messages). This solution aims to reduce the burden placed on individual users to discern whether a text message is legitimate or not. By implementing our robust model, we are able to support both individuals and organizations in filtering our potential scams efficiently and reliably.

GOALS

Our primary goal is to develop a highly accurate, precise, and reliable spam detection model that enhances customer experience by accurately filtering spam with minimal disruptions to legitimate messages. Our focus on precision, specificity, and overall model effectiveness directly supports customer satisfaction by ensuring only valuable messages reach them. The KPIs that will measure our success includes:

1. High Precision (> 95%): High precision minimizes false positives, meaning legitimate messages are rarely flagged as spam. This preserves the flow of essential communication for our customers. The accuracy in minimizing false positives is crucial for maintaining customer trust and satisfaction.
2. High Specificity (>97%): Achieving high specificity allows our model to consistently identify ham messages as legitimate. This KPI helps ensure minimal interference with normal user communication, reinforcing a seamless user experience and aligning with our goal to offer a more reliable and non-intrusive service.
3. F1 Score (>90%): The F1 score balances precision and recall, providing a comprehensive measure of the model's effectiveness in capturing spam while minimizing errors. A high F1 score demonstrates the model's reliability in practical scenarios, supporting customer satisfaction and business reliability.

Additional KPIs such as ROC-AUC, Confusion Matrix, Recall, and Accuracy will further gauge the model's performance. The ROC-AUC score will visualize the model's true-positive rates and the Confusion Matrix will provide granular insight into specific performances of each classification. We will validate these KPIs by using real-world spam messages as a test to verify the model meets customer needs.

DATASET ANALYSIS &EXPLANATION

The data set contains SMS messages labeled as spam or ham(non-spam), sourced from real-world examples in text classification. We selected this dataset for text classification task training in our NLP model because it provides realistic and diverse examples of spam. From this dataset, we can better train our model to accurately and precisely identify if a message is spam or ham.

In terms of bias in the ratio of spam to ham, our dataset contains a 87:13 split which corresponds with 86:14 split in a recently published paper on “Securing Mobile Message Communications”. We do not predict our model favoring either spam or ham. The model may perform less well when applied to messages outside of this context of the US English language since the messages in the dataset are pulled from North America.

Pros:

- Pre-labeled: Ideal for supervised learning, labels are already assigned, enabling faster model training.
- Balanced Lengths: SMS messages are brief which aligns with NLP and LLM models using text classification algorithms.
- Relevance to the real world: Aims to address spam filtering challenges, applicable to commercial SMS filtering systems.

Cons:

- Outdated Data: The model's generalizability may be limited because the nature of spam messages may have changed since the dataset was collected (e.g., new sorts of scams).

METHODS

1. Text pre-processing
    a. Lowercasing: Standardize text for uniformity.
    b. Punctuation and Stop-word Removal: Removes non-essential words (e.g., "a," "the," "and") that do not contribute to distinguishing spam from ham.
    c. Tokenization: Breaks down text into words, allowing word-based analysis.
    d. Stemming and Lemmatization: Reduces words to their root forms (e.g., "running" to "run"), improving word consistency across texts.
2. Feature Extractions: Process of pulling specific information from raw text to create numerical representation such as numeric values and vectors. Enables our model to see patterns and associations to distinguish neutral words (ham texts) from spam words.
    a. Bag of words: Counts word occurrences, helping the model focus on commonly repeated words typical of spam.
    b. TF-IDF (Term Frequency-Inverse Document Frequency): Weights words based on their rarity across all texts, emphasizing unusual words that may signal spam.
    c. N-grams (Bigrams, Trigrams): Captures key word pairs or triplets, such as "Click here," indicative of spam.
    d. Word Embedding: Transforms words into vectors that represent semantic relationships. May leverage pre-trained embeddings if available to save on training time.
3. Feature Engineering: To enhance model precision, we'll extract specific features that improve spam detection accuracy:
    a. Special character counts: Spam often includes symbols like "!!!" to draw attention.
    b. Capitalization Ratio: Excessive capitalization can signal urgency, a common spam tactic.
    c. Text length: Extreme message lengths often correlate with spam.
    d. Presence of URL/Links: Many spam messages contain suspicious URLs.
    e. Sentiment Analysis Score: Detects overly positive or urgent tones common in scams.

MODEL DEVELOPMENT & SELECTION

Our main approach will focus on NLP classification models, and we will experiment with methods ranging from simpler algorithms, like k-Nearest Neighbors (kNN), to more advanced machine learning models. Additionally, we're considering neural networks if further tests reveal they improve model accuracy or better capture the semantic nuances of spam patterns.

By following this structured methodology, we aim to build an effective spam detection model that supports our business objectives to enhance customer experience and brand reputation.

CONCLUSION

This project aims to develop a high-precision, reliable spam detecting model that protects users from phishing and spam messages while enhancing overall customer satisfaction. By reducing user engagement in the filtering process, we streamline communication and minimize potential risks associated with fraud and scams, safeguarding user's private information and security.

Our model's focus on key performance indicators like precision, specificity, and the F1 score ensures that an approach that is technically sound and scalable while aligning with the company's business objective of delivering a more secure SMS service. As the model evolves, it will allow us to offer customers a safer, distraction-free communication experience, adding measurable value to our brand.

From a business perspective, this initiative has strong potential to become a profit generator by reducing costs associated with fraud mitigation and by increasing brand recognition and trust. Differentiating from our competitors will contribute to customer loyalty and better position us in the marketplace as a trusted provider of secure communications.

As we continue to refine the model based on real-world data and adapt to emerging trends, this project will remain flexible, with ongoing enhancements that ensures we meet current and future requirements. This proposal presents a solution that not only mitigates spam and scam risks but also fortifies our business as a trusted and innovative service provider.