Sistema de Classificação de SMS para Identificação de Spam utilizando o Classificador Ingênuo de Bayes

1st Amanda Cristina Fernandes M. de Lima
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
acfml@cin.ufpe.br

2nd Maria Letícia do N. Gaspar
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mlng@cin.ufpe.br

3rd Mariana Melo dos Santos
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
mms11@cin.ufpe.br

4th Victória Barbosa C. Figueiredo
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
vbcf@cin.ufpe.br

***Abstract*—Este projeto visa utilizar o classificador probabilístico Ingênuo de Bayes para identificação eficiente de Spam, distinguindo mensagens indesejadas de conteúdos legítimos através da análise de atributos textuais e estruturais.**

***Index Terms*—SMS, spam, Naïve Bayes, classificador, identificação**

I. Introdução

Spam é uma prática de enviar SMS's, e-mails ou mensagens sem o consentimento do usuário na internet. Essas mensagens possuem conteúdos indesejados ou perigosos para o usuário como: propagandas, vírus, fake news ou estelionato.Segundo a revista Exame, 66% dos brasileiros recebem spam por SMS toda semana, ou seja, infelizmente essa é uma prática extremamente comum. Assim, é preciso encontrar uma forma de detectar se os SMS's são ou não spam para que o usuário não receba diversas mensagens desagradáveis ou perigosas e também não deixe de receber SMS's de seu interesse.

II. Objetivos

O principal objetivo desse projeto é desenvolver um modelo que possibilite classificar um SMS como spam ou não através do classificador ingênuo de Bayes. Além disso, será necessário realizar uma análise exploratória dos dados para compreender melhor os padrões presentes nos dados. Por fim, o desempenho do modelo será avaliado usando métricas de avaliação padrão.

III. Justificativa

A utilização de SMS's continua sendo uma das principais formas de comunicação digital e é amplamente utilizada em diversos contextos, tanto pessoais quanto profissionais. No entanto, todos os dias, os mensagens de texto de spam tornam-se mais inteligentes e mais sábios, apresentando um conteúdo melhor e desenvolvendo novos métodos capazes de ultrapassar os filtros de spam empregados atualmente.
Nesse contexto, existe a necessidade de desenvolver um modelo de filtragem baseado no algoritmo *Naïve Bayes* tanto para evitar que mensagens legítimas sejam classificados erroneamente como spam, quanto para que a caixa de entrada de usuários não seja poluída de conteúdos indesejados e possivelmente perigosos.

IV. Base de Dados

A base de dados de spam (*spambase*) utilizada no desenvolvimento do modelo foi fornecida pela *UCI Machine Learning Repository* e consiste em 5.572 mensagens de texto (SMS), que possuem uma estrutura textual semelhante à e-mails com um número de caracteres menor. Além disso, 4.825 ou $86,6\%$ dessas mensagens são conteúdos legítimos (*ham*) com 4.516 delas contendo textos únicos e sendo "*Sorry, I'll call back later*" a sentença mais repetida. Consequentemente, 747 ou $13,4\%$ dos SMS's na base de dados são mensagens ilegítimas (spam), sendo 653 delas conteúdos únicos e "*Please call our customer service representativ...*" a frase que mais se repete.
Esse repertório nos permite analisar, filtrar e classificar qualquer tipo de texto como spam ou não spam (ham), além de facilitar o desenvolvimento do modelo por possuir mensagens de tamanho mais reduzido, tornando assim a análise mais simples, rápida e eficaz.

V. Análise Exploratória dos Dados

A. Visão Inicial dos dados

Após fazermos uma visualização inicial da nossa base de dados, podemos perceber que ela possui 5 colunas:

- v1: indica se a mensagem de texto naquela coluna é um spam ou não (ham);
- v2: indica o conteúdo da mensagem de texto;
- unnamed 2, unnamed 3, unnamed 4: não apresentam informações relevantes para o desenvolvimento do nosso modelo

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
847	ham	I am in office:)whats the matter..msg me now.i...	NaN	NaN	NaN
3009	ham	Imagine Life WITHOUT ME... see.. How fast u ar...	NaN	NaN	NaN
17	ham	Eh u remember how 2 spell his name... Yes i di...	NaN	NaN	NaN
2070	spam	Sexy Singles are waiting for you! Text your AG...	NaN	NaN	NaN
611	ham	Its a valentine game. . . Send dis msg to all ...	NaN	NaN	NaN

Fig. 1. Amostra de 5 linhas da spambase

B. Remoção e Renomeação de Colunas

Tendo em vista que as últimas colunas não serão utilizadas no desenvolvimento do sistema de classificação, iremos removê-las do repertório. Além disso, para facilitar a visualização dos dados, renomearemos as colunas v1 e v2 para 'categoria' e 'mensagem', respectivamente.

	categoria	mensagem
2993	ham	No idea, I guess we'll work that out an hour a...
2910	ham	Sorry,in meeting I'll call later
3389	spam	Please CALL 08712402972 immediately as there i...
1232	ham	1's finish meeting call me.
5155	ham	MY NEW YEARS EVE WAS OK. I WENT TO A PARTY WIT...

Fig. 2. Amostra de 5 linhas da spambase modificada

C. Agrupamento dos Dados por Categoria

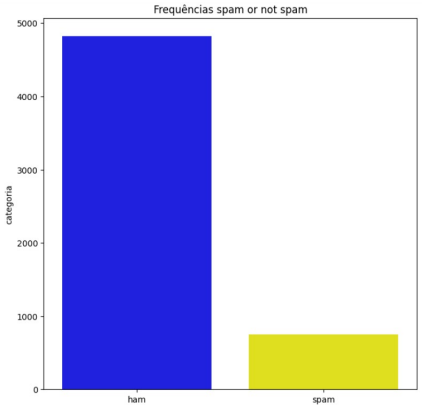

Fig. 3. Gráfico de frequência: base de dados agrupada

Para melhor visualização e, também, para facilitar a classificação de spam pelo nosso modelo, agrupamos os dados da base em dois grupos: ham (mensagens legítimas) e spam (mensagens ilegítimas).

	mensagem			
	count	unique	top	freq
categoria				
ham	4825	4516	Sorry, I'll call later	30
spam	747	653	Please call our customer service representativ...	4

Fig. 4. Base de dados agrupada

Podemos perceber que a sentença que mais se repete do grupo ham ("*Sorry, I'll call later*") aparece no repertório numa frequência de 30 vezes, enquanto a sentença que mais se repete do grupo spam ("*Please call our customer service representativ...*") aparece em uma frequência de 4 vezes.

D. Análise do Tamanho dos SMS por grupo

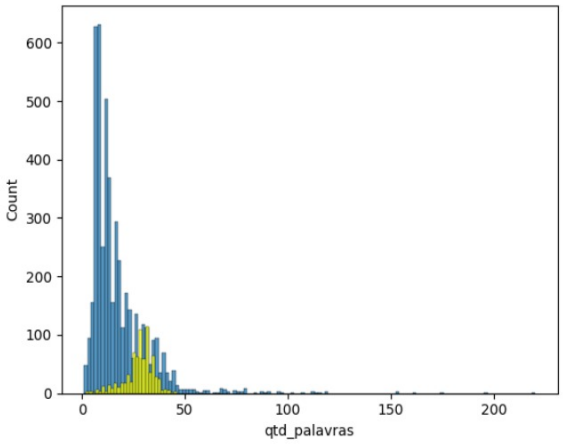

Fig. 5. Quantidade de mensagens por número de palavras

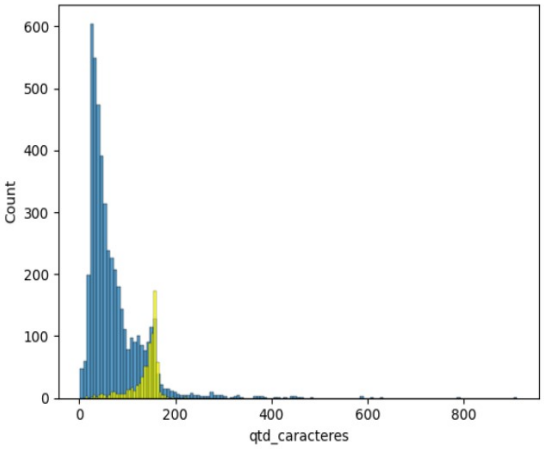

Fig. 6. Quantidade de mensagens por número de caracteres

A partir desses gráficos, podemos perceber que os SMS's que fazem parte do grupo spam tentem a ser maiores (conter mais palavras/caracteres) que os SMS's do grupo ham.

VI. Pré-processamento dos dados

O processo de pré-processamento de dados para a classificação de SMS como spam ou não spam é fundamental para garantir a eficácia do modelo. Esse processo envolve várias etapas essenciais:

- Conversão para letras minúsculas: Inicialmente, todas as letras do texto são convertidas para letras minúsculas. Isso ajuda a evitar que o modelo considere palavras em maiúsculas e minúsculas como diferentes, contribuindo na consistência da análise.
- Remoção de pontuação: A pontuação, como vírgulas, pontos de exclamação e interrogação, é removida do texto. Isso simplifica o texto e evita que o modelo considere elementos não informativos.
- Remoção de stopwords: Stopwords são palavras comuns que geralmente não contribuem para análise, como srtigos e preposições. Essas palavras são removidas para reduzir o ruído nos dados.
- Tokenização: O texto é dividido em tokens, que são as unidades individuais, geralmente palavras. Isso facilita a análise subsequente, pois cada palavra se torna uma unidade separada.
- Vetorização: Para que o modelo de classificação de spam possa processar os dados de texto, as palavras precisam ser convertidas em representações numéricas. Isso é feito por meio da vetorização, onde cada palavra é mapeada para um número inteiro único.
- Ponderação TF-IDF (Term Frequency-Inverse Document Frequency): A ponderação TF-IDF é aplicada aos vetores de palavras para atribuir valores que refletem a importância relativa de cada palavra em relação ao conjunto de documentos. Isso ajuda a destacar palavras-chave que podem ser indicativas de spam, pois são menos comuns em mensagens legítimas e mais comuns em mensagens de spam.

VII. Classificador Ingênuo de Bayes

Nesse projeto,utilizaremos o algoritmo ingênuo de Bayes que é um algoritmo de aprendizagem de máquina supervisionado. Esse classificador foi criado por Thomas Bayes que nasceu em Londres e foi um pastor presbiteriano e matemático. Além disso, foi eleito membro da Royal Society em 1742.
O algoritmo de Bayes é aplicado, por exemplo, em classificadores de sentimentos e na área de saúde para determinar se há ou não a ocorrência de uma doença. Só não é muito indicado para casos muito complexos que envolvam uma quantidade muito grande de dados ou variáveis numéricas.

Esse modelo considera as features (variáveis) independentes entre si e possui uma base matemática relativamente simples e também um bom desempenho em comparação com os demais classificadores.

Uma vantagem desse algoritmo é que ele depende de poucas dados para ter uma boa acurácia. O algoritmo funciona da seguinte maneira: para calcular a predição, o algoritmo irá definir uma tabela de probabilidades, em que consta a frequência dos preditores com relação às variáveis de saída. Então, o cálculo final leva em conta a probabilidade máxima para oferecer uma solução.
O algoritmo utiliza o Teorema de Bayes (cálculo da probabilidade de um evento dado que outro evento já ocorreu) que possui a fórmula:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{1}$$

- P(A) : Probabilidade do evento A ocorrer
- P(B) : Probabilidade do evento B ocorrer.
- $P(B|A)$: Probabilidade de B ocorrer sabendo que A já ocorreu.
- $P(A|B)$: Probabilidade de A ocorrer sabendo que B já ocorreu.

Considerando o evento A como a ocorrência de spam e o evento B como a ocorrência de uma determinada palavra, teremos a aplicação desse conceito na classificação de e-mails como spam e não spam.
A partir dessa métrica, é possível desenvolver um filtro capaz de detectar os SMS's indesejados e separá-los dos legítimos. Isso é feito por meio de uma medição do quanto de lixo eletrônico cada SMS está enviando, ou seja, é calculada a probabilidade de cada palavra no SMS ser spam e, então, tais probabilidades são multiplicadas pela quantidades de ocorrências de cada termo no SMS, obtendo-se a métrica geral de spam capaz de classificar tais mensagens.
Foi escolhida a classe MultinomialNB para realizar a distinção de SMSs spam e não spam, pois é um algoritmo de aprendizado de máquina amplamente utilizado na área de processamento de linguagem natural e classificação de texto. Ele é especialmente eficaz quando se lida com dados textuais que envolvem contagens de ocorrências de palavras ou termos.
O funcionamento do MultinomialNB envolve a estimativa das probabilidades condicionais de cada classe com base nas frequências dos termos nas amostras de treinamento. Essas probabilidades são então usadas para classificar novos documentos ou textos desconhecidos.

VIII. Análise dos Resultados

A partir do desenvolvimento que obtivemos no nosso sistema de classificação de spam em SMS utilizando o Classificador Ingênuo de Bayes (CIB) e da análise feita acerca da base de dados utilizada, é possível perceber que o modelo obteve 96% de acurácia.

Como é possível observar na figura 7, o modelo identificou e classificou corretamente 100% dos SMS's do grupo spam. Entretando, esse percentual no caso do grupo ham foi de 95%, indicando a existência de casos em que o sistema classificou

	precision	recall	f1-score	support
0	0.95	1.00	0.98	477
1	1.00	0.72	0.83	81
accuracy			0.96	558
macro avg	0.98	0.86	0.91	558
weighted avg	0.96	0.96	0.96	558

Fig. 7. Acurácia do sistema de classificação de spam

incorretamente um SMS legítimo como spam.
Visualizando a classificação dos SMSs do conjunto de dados teste na matriz de confusão, tem-se que:

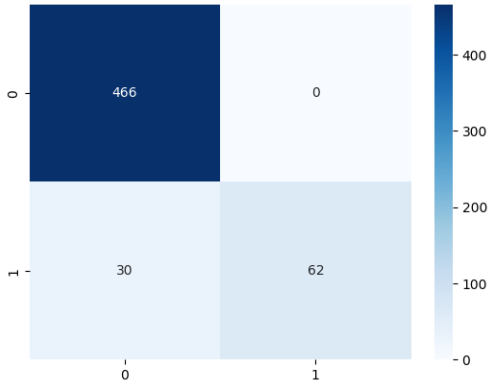

Fig. 8. Matriz de confusão

- (0,0) - verdadeiro positivo: SMS que não são spam e foram classificados corretamente
- (0,1) - falso negativo: SMS que são spam e foram classificados como não spam
- (1,0) - falso positivo: SMS que não são spam e foram classificados como spam
- (1,1) - verdadeiro negativo: SMS que são spam e foram classificados corretamente

IX. Conclusão

Em resumo, nosso sistema de classificação de SMS usando o Classificador Ingênuo de Bayes atingiu uma impressionante taxa de acurácia de 96%. Isso significa que o modelo mostrou-se bastante eficaz na identificação de mensagens de spam, tornando a experiência do usuário mais segura. No entanto, é importante continuarmos melhorando e atualizando nosso sistema para lidar com as novas táticas dos spammers. No geral, este projeto mostra como o uso do *Machine Learning* pode ajudar a proteger os usuários e melhorar a qualidade das comunicações digitais.

References

[1] M. Paul, "Probabilidade: Aplicações à Estatítica". 2 Edição. livros Técnicos e Científicos Editora.
[2] https://archive.ics.uci.edu/dataset/228/sms+spam+collection
[3] Spambase [https://archive.ics.uci.edu/dataset/94/spambase]
[4] https://kinsta.com/pt/blog/por-que-meus-e-mails-estao-indo-para-o-spam/
[5] https://www.spamlaws.com/spam-stats.html
[6] https://www.bbc.com/portuguese/internacional-59701523