

USO DE VISUALIZAÇÃO DA INFORMAÇÃO NA ANÁLISE DOS DADOS GERADOS PELA CLASSIFICAÇÃO DE E-MAILS POR MEIO DE CÁLCULOS ESTATÍSTICOS BAYESIANOS

USE OF INFORMATION VISUALIZATION IN THE ANALYSIS OF DATA GENERATED BY THE CLASSIFICATION OF EMAILS THROUGH BAYESIAN STATISTICAL CALCULATIONS

Gabriela Guilherme de Almeida¹; Robson Augusto Siscoutto²; Ronaldo Toshiaki Oikawa³

Universidade do Oeste Paulista - Faculdade de Informática de Presidente Prudente^{1,2,3}
gabih.g.a@hotmail.com¹; robson@unoeste.br²; oikawa@unoeste.br³

RESUMO - Juntamente com a popularização e facilidade de acesso da internet, surgiram alguns problemas e o spam é um dos mais difíceis de ser resolvido. Sendo assim, o estudo de tais mensagens se torna indispensável para a evolução de ferramentas de controle. O uso de métodos estatísticos bayesianos contribui para a identificação e controle do recebimento de spam, entretanto, tais filtros geram uma vasta quantidade de dados estatísticos de alta complexidade para serem analisados. Dessa forma, o trabalho proposto neste artigo, visa estudo e a aplicabilidade de cálculos estatísticos bayesianos no reconhecimento de spams em e-mails e a visualização dos dados gerados por esta análise em representações visuais 3D organizadas de maneira intuitiva e interativa, facilitando o processo de tomada de decisão.

Palavras-chave: Spam; Filtro Bayesiano; Visualização da Informação; Realidade Virtual.

ABSTRACT - Along with the popularity and ease of access from the internet, there were some problems and spam is one of the most difficult to be solved. Thus, the study of such messages becomes essential for the evolution of control tools. The use of Bayesian statistical methods contributes to the identification and control of receiving spam, however, such filters generate vast amounts of highly complex statistical data to be analyzed. In this way, the work proposed in this article aims to study and the applicability of Bayesian statistical calculations in the recognition of spam in email and display the data generated by this analysis in 3D visual representations organized intuitively and interactively, facilitating the decision-making process.

Keywords: Spam; Bayesian Spam Filtering; Information Visualization; Virtual Reality.

Recebido em: 15/04/2015
Revisado em: 10/10/2015
Aprovado em: 07/11/2015

1 INTRODUÇÃO

Com o avanço da internet nos últimos anos, surgiram novos meios de se fazer publicidade e um deles é o spam, que é definido como “toda mensagem eletrônica não desejada pelo destinatário” (TAVEIRA, 2008, p.1). Muitas vezes, ao clicar em um link contido em tais mensagens, são instalados vírus no computador do usuário, sem que ele perceba, por onde as informações são roubadas ou apagadas.

Alguns spammers (responsáveis pelo envio de spams) usam contas gratuitas em provedores legítimos para disseminar mensagens indesejadas, sendo assim, alguns provedores limitam a quantidade de mensagens enviadas por dia, o que não é eficiente devido a facilidade de se criar vários e-mails.

Outro método de se verificar se um e-mail é ou não spam pode ser feito pelo uso de filtros por métodos estatísticos, que visam identificar e controlar o recebimento de spam. Tais métodos

geram uma base de dados estatísticos, que precisa ser analisada pelo gestor e que desprende muito tempo.

A quantidade de dados nessas bases gera informações estatísticas que dificultam o processo de análise, compreensão e utilização dos dados, bem como a tomada de decisão. Muitas dessas informações são irrelevantes para o usuário, ou as informações consideradas úteis podem simplesmente ser perdidas, devido aos usuários não conhecerem o relacionamento entre os dados.

Uma abordagem que pode facilitar esse processo de análise consiste na aplicação de técnicas de Visualização da Informação, que estudam formas de transformar dados abstratos em representações visuais, de forma a facilitar o seu entendimento e/ou ajudar na descoberta de novas informações contidas nestes dados (NASCIMENTO, 2005). Tais representações podem ser distribuídas espacialmente em forma unidimensional, bidimensional e/ou tridimensional, que são definidas de acordo com a dimensão do espaço onde os elementos geométricos utilizados estejam situados (FREITAS et al., 2001). A visualização da informação pode ser realizada por meio de ambientes de Realidade Virtual e Aumentada, facilitando a análise e a compreensão dos dados, já que os mesmos podem ser dispostos de maneira intuitiva e interativa. Mesmo que a visualização da informação contribua para facilitar o processo de análise dos cálculos de spam, ainda é pouco utilizada.

Apesar dos estudos na área, ainda há falhas nos métodos de classificação por diversos autores e algumas mensagens legítimas são taxadas como spam (falsos positivos) ou mensagens não legítimas classificadas como não spam (falsos negativos). Diante disso, problemas como sobrecarga dos servidores com e-mails falsos negativos, ilusão por parte do usuário de que sua caixa de mensagem está segura ou livre de spam, execução de códigos maliciosos e dificuldade na tomada de decisão devido à análise manual de dados estatísticos podem ocorrer.

Outro ponto está na eficiência de tais métodos estatísticos que ainda são discutíveis, principalmente, pela quantidade diária de mensagens enviadas. Portanto, ainda é necessário que filtros mais eficientes sejam desenvolvidos ou estudados, visando buscar o equilíbrio entre filtrar mensagens indesejadas sem que mensagens "verdadeiras" sejam bloqueadas. Além disso, também há a necessidade de facilitar a análise dos dados gerados pela classificação, proporcionando a

tomada de decisão mais consciente por parte do gestor.

Sendo assim, o objetivo do presente trabalho é desenvolver uma ferramenta que proporciona a visualização de dados em gráficos 3D gerados pela análise e classificação de e-mails fazendo uso de cálculos estatísticos bayesianos.

Este artigo está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados; a seção 3 apresenta a fundamentação do trabalho, os recursos, materiais e métodos utilizados; a seção 4 apresenta o escopo do projeto (funcionalidades, diagramas); a seção 5 apresenta a arquitetura e visão geral da aplicação; a seção 6 apresenta os resultados obtidos e a seção 7 as considerações finais sobre o trabalho.

2 TRABALHOS RELACIONADOS

Esta seção tem como objetivo apresentar os trabalhos relacionados e seus respectivos resultados.

2.1 Filtro de Conteúdo para Sistemas SMS Baseado em Classificador Bayesiano e Agrupamento por Palavras

Belém (2009) propõe um filtro de conteúdo para sistemas SMS baseado em Classificador Bayesiano e agrupamento de palavras. Segundo ele, o filtro foi implementado em uma entidade ESME (*External Short Message Entity*) responsável por entregar as mensagens enviadas pelos provedores de conteúdo à SMSC (*Short Message Service Center*).

As palavras são armazenadas em um vetor e este é percorrido sequencialmente pegando de uma a cinco palavras dependendo da escolha do usuário e tais combinações são armazenadas. Dessa forma, cada posição do vetor é um conjunto de palavras e um atributo associado.

Inicialmente obteve-se uma quantidade de mensagens de uma ESME e essas mensagens foram classificadas manualmente. Segundo Belém (2009), "foi estabelecida uma probabilidade de cada uma dessas palavras e grupos de palavras aparecerem em uma mensagem spam e não spam, por meio da fórmula de Bayes".

Segundo Belém (2009): "O filtro estará em constante atualização, a partir da chegada de novas mensagens, as probabilidades de cada uma das palavras são atualizadas".

O desenvolvimento do algoritmo foi dividido em três módulos: treinamento, classificação e aprendizado. No primeiro módulo, foram inseridas mensagens previamente

classificadas como spam e não spam, no segundo, o algoritmo recebe uma mensagem e a classifica como spam ou não spam e já o terceiro, os usuários que possuem o sistema operacional Android puderam classificar as mensagens por meio de um aplicativo.

Foram realizados experimentos após o treinamento e a Tabela 1 apresenta os resultados para a classificação das palavras e atributo.

	Uma palavra e atributos	Doas palavras atributos	Três palavras e atributos	Quatro palavras e atributos	Cinco palavras e atributos
Tempo em segundos do teste	53,815	99,041	131,117	148,866	181,827
Tempo gasto por mensagem em segundos	0,006	0,011	0,015	0,017	0,020
Número de mensagens	8.687	8.687	8.687	8.687	8.687
Quantidade de mensagens spam	8.008	8.008	8.008	8.008	8.008
Quantidade de mensagens não spam	636	636	636	636	636
Falso positivo	29	122	133	139	140
Falso negativo	380	64	42	42	42
Grupos utilizados	69.515	386.910	717.550	1.073.368	1.405.285
Memória (MB)	0,49	4,38	11,05	19,95	30,32

Tabela 1. Resultados para classificação palavra e atributos (BELÉM, 2009).

A partir dos resultados obtidos, conclui-se que a medida que aumenta-se a quantidade de palavras, a quantidade de falsos positivos também aumenta, entretanto, o número de falsos negativos diminui. Percebe-se também que o filtro é mais eficiente quando se utiliza grupo de até três palavras. Vale ressaltar que entre a utilização do grupo de uma palavra e do grupo de até duas palavras, o percentual aumentou de forma significativa e que quando foi utilizado o grupo de até quatro e cinco, o desempenho teve uma ligeira queda.

2.2 Mecanismo Anti-Spam Baseado em Autenticação e Reputação

Taveira (2008) propõe um filtro baseado em autenticação e reputação com o intuito de reduzir os falsos positivos. A principal ideia deste método é levar em consideração o histórico de comportamento, ou seja, se um usuário enviar uma grande quantidade de mensagens legítimas, ele possui uma boa reputação e, portanto, as mensagens que este envia tem uma maior probabilidade de serem legítimas.

Primeiramente, procura-se identificar de forma exata e sem utilizar informações pessoais do usuário que enviou uma determinada mensagem. Para isso, a autenticação é feita por meio de pseudônimos.

Feita a identificação do usuário, é possível monitorar seu histórico de comportamento e este é determinado por meio de diversos servidores de e-

mail que trocam informações. Nesta troca é utilizado o mecanismo de reputação para avaliar os servidores que fornecem informações e conseqüentemente seu grau de confiança. Taveira (2008) afirma que os resultados obtidos com o mecanismo proposto são satisfatórios, pois permite a redução de falsos positivos sem prejudicar a taxa de falsos negativos e é eficiente contra as tentativas elaboradas pelos spammers para burlar o filtro.

A reputação dos servidores é determinada a partir da quantidade de mensagens legítimas e spam que este envia. Para classificar tais mensagens, são utilizados filtros convencionais (listas brancas e negras, filtros bayesianos e entre outros) como mecanismo auxiliar. Servidores podem apresentar um bom comportamento por um período de tempo e em seguida passam a enviar spams, portanto, para que isso seja barrado pelo filtro, o cálculo da reputação é feito “de acordo com a média móvel exponencial, que considera pesos maiores aos valores observados mais recentemente e os pesos dos elementos mais antigos decrescem exponencialmente” (TAVEIRA, 2008, p.39). Sendo assim, as mensagens mais antigas tem um peso menor no cálculo.

Os resultados foram colhidos por meio de simulações e Taveira (2008) afirma que “as simulações mostram a eficiência do mecanismo proposto, que reduz em 537 vezes a taxa de falsos positivos quando a taxa de falsos positivos do mecanismo auxiliar é de 10%, ou seja, uma redução de 99,81%” (TAVEIRA, 2008, p.79).

2.3 Multidimensional Information Visualization Using Augmented Reality

Meiguins et al. (2006) propõe um protótipo para a visualização de informações em um ambiente de RA que permite a filtragem dos dados, seleção dos atributos, zoom semântico e entre outros. A interação com o usuário ocorre de forma direta por meio de filtros e manipulação de objetos 3D e também é possível a manipulação de objetos virtuais.

Um pré-requisito importante para a implementação da técnica de visualização é o tratamento da base de dados. Durante esta fase, é recomendada a utilização de técnicas de seleção de atributos para reduzir o volume de dados existente, utilizando apenas os atributos que contribuem para uma boa análise dos dados.

Para o teste do protótipo foi utilizada uma base de domínio público que contém informações sobre automóveis brasileiros. Depois da preparação dos dados foi gerada uma tabela contendo 41 linhas

e 10 colunas. Uma parte da tabela é mostrada na Tabela 2.

MAKE	MODEL	FUEL	YEAR	COLOR	PRICE	DOOR	TYPE	POWER	RPM
STRING	STRING	CADEIA	INT	STRING	REAL	INT	STRING	INT	INT
Fiat	Uno	gasolina	1993	prata	6900	4	hatch	55	4000
Kia	Sportage	diesel	1995	prata	40000	4	pickup	120	4000
volkswagen	Parati	gasolina	1993	branco	9400	2	wagon	99	5600

Tabela 2. Dados dos carros brasileiros (MEIGUINS et al., 2006).

No modelo de gráfico apresentado, os objetos 3D são posicionados de acordo com o eixo x, y e z. Cada objeto visualizado tem características específicas de cor, forma e tamanho que representam, diretamente, os itens de dados valores. Isto caracteriza a técnica implementada como multidimensional (informações de eixo x, y e z, cor, forma e tamanho). Nas configurações, o eixo x é definido para o atributo ano, o eixo y para o preço do automóvel, o eixo z para o combustível do automóvel, a cor representa a marca, o tamanho o número de portas e a forma o tipo de carro.

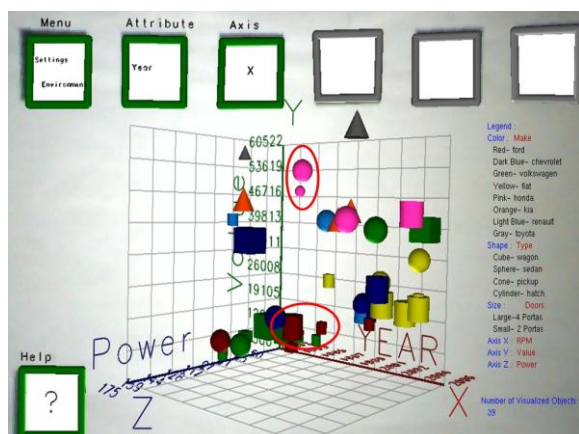


Figura 1. Configurações iniciais (MEIGUINS et al., 2006).

Na figura 1 os objetos circulosados em vermelho podem ser interpretados pela legenda do lado do gráfico. A esfera rosa maior representa o Sedan - Honda - 4 portas e a menor é Sedan - Honda - 2 portas, já o cilindro vermelho maior representa o Hatch - Ford - 4 portas e o menor Hatch - Ford - 2 portas.

A interação do usuário com a interface gráfica do protótipo acontece de forma direta por meio da oclusão do marcador, onde o controle virtual desejado e projetado para uma determinada ação a ser executada. A interação por oclusão consiste em obstruir a captura de um marcador pela webcam.

O usuário tem a opção de fazer consultas dinâmicas, que permitem aos usuários realizar consultas em uma base de dados sem a necessidade de linhas de comando, utilizando apenas componentes da interface gráfica.

Os atributos discretos (ou nominais) são geralmente utilizados como opções de configuração, como cor, forma e tamanho. A Figura 2 ilustra a filtragem de apenas uma forma, também é possível realizar a remoção de atributos desejados do gráfico.

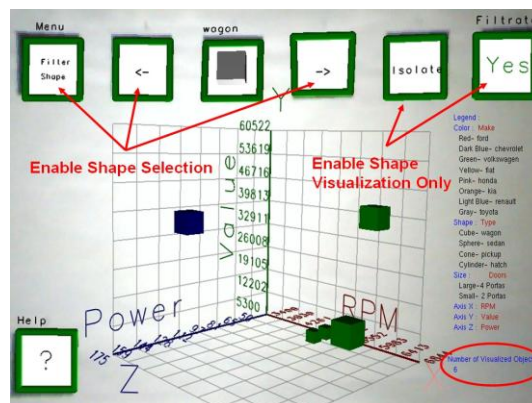


Figura 2. Filtragem de atributos discretos isolando uma forma (MEIGUINS et al., 2006)

O protótipo permite visualizar as informações de cada objeto, além de permitir o zoom semântico que possibilita que o usuário visualize áreas específicas do gráfico.

2.4 Uso de realidade aumentada em ambientes virtuais de visualização de dados

Buk (2006) propõe uma ferramenta de visualização de dados iterativa utilizando Realidade Virtual e Realidade Aumentada. O DATAVIS-AR, permite a interação com o usuário por meio de gráficos configuráveis e reutilizáveis.

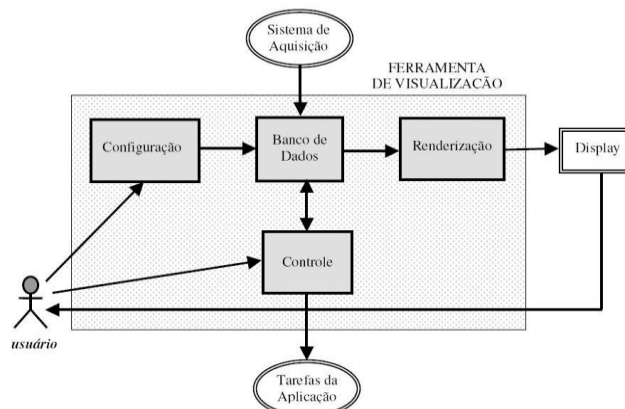


Figura 3. Funcionamento do DATAVIS-AR (BUK, 2006, p.59)

Na Figura 3, observa-se o esquema de funcionamento da ferramenta. O sistema de aquisição alimenta continuamente o banco de dados. O módulo de controle filtra os dados de acordo com o que foi especificado pelo usuário e é por meio desse módulo que o usuário é capaz de interagir com o gráfico e essa interação dispara uma tarefa da aplicação.

O módulo de configuração é o que permite o usuário ajustar o sistema conforme sua necessidade, por meio dele, é permitido o ajuste do banco de dados estabelecendo restrições e ajustes nos fatores de decisão.

O módulo de renderização monitora o banco de dados e prepara os gráficos que os usuários irão visualizar.



Figura 4. Exemplo de uso do DATAVIS-AR (BUK, 2006, p. 62)

No exemplo apresentado na Figura 4, foi construído um tabuleiro com partes de gráficos, o que permite a análise específica de atributos. As placas dos gráficos, seus respectivos marcadores e os gráficos correspondentes foram cadastrados de antemão no ARToolKit, sendo assim, quando a placa é colocada em frente a uma webcam, a ferramenta reconhece e insere o gráfico que a corresponde.



Figura 5. Exemplo do uso do DATAVIS-AR (BUK, 2006, p. 61).

Na Figura 5 é apresentado outro exemplo, em que se observa a chamada de um gráfico em VRML que aparece na mão do usuário. Dessa forma, o usuário pode girar o gráfico e assim, analisa-lo de forma mais precisa.

3 FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão detalhados os principais conceitos teóricos utilizados como base para o desenvolvimento deste.

3.1 Listas brancas e negras

De acordo com Taveira (2008, p. 12), os primeiros sistemas AntiSpam surgiram baseavam-se na utilização de listas negras, que são listas com endereços de origem ou endereços IP de remetentes que reconhecidamente são fontes de spam. Nas listas brancas, são colocados endereços de pessoas ou servidores confiáveis. Dessa forma, quando uma mensagem é recebida, as duas listas são consultadas (TAVEIRA, 2008).

Sendo assim, se o endereço de origem estiver na lista negra, este será classificado diretamente como spam e se estiver na lista branca, vai para a caixa de entrada. Caso o endereço remetente não estiver em nenhuma das duas listas, são usados filtros AntiSpam disponíveis.

Primeiramente, as listas eram feitas de forma manual pelos usuários, mas isto gerava um grande incômodo para os gerentes destas listas. Segundo Taveira (2008), foi criado um sistema centralizado e uma organização que controla os endereços colocados ou retirados destas listas.

3.2 Visualização da informação

As técnicas de Visualização da informação têm como objetivo transformar um conjunto de dados brutos, nos quais se procura informação em uma forma visual compreensível e manipulável pelo usuário (ZORZAL; CARDOSO; KIRNER, 2009, p.22). Segundo ele, a visualização da informação possui três etapas:

- Transformação dos dados: que consiste na transformação dos dados brutos e a sua devida organização em uma estrutura lógica.
- Mapeamento visual: que significa construir uma estrutura visual para os dados que foram organizados logicamente. A estrutura visual se divide em três partes: substrato espacial, que é o espaço para a visualização; marcas visuais, que são os símbolos gráficos usados para representar os dados; e propriedades gráficas das marcas, que são os atributos que caracterizam as marcas.

- Transformações visuais: é possível que o usuário interaja com a representação gráfica.

Buk (2006) afirma que a visualização da informação pode ser classificada em unidimensional, bidimensional, tridimensional e multidimensional em relação as informações apresentadas.

A visualização unidimensional é baseada em um elemento único sequencialmente organizado. Já a visualização bidimensional “trabalha dois aspectos: a posição que o elemento está e um atributo. O atributo pode ser a altura, cor, largura” (BUK, 2006, p. 27). A tridimensional utiliza os aspectos da bidimensional mais o volume do objeto. A multidimensional se dá por meio de adição de atributos (como tonalidade, forma, cor e entre outros) em gráficos bidimensionais.

3.2 Realidade Virtual e Realidade Aumentada

A realidade virtual, também chamada de RV, segundo Cardoso et al. (2007), tem como função criar para o usuário um ambiente virtual no qual ele possa interagir com os objetos presentes de maneira intuitiva e natural. Ou seja, quando o usuário estiver imerso neste ambiente tridimensional projetado pelo computador ele deve tocar e modificar os objetos que compõem este ambiente.

A Realidade Aumentada, chamada de RA, tem como função, associar dados computacionais ao mundo real e permitir ao usuário a interação com tais dados. Segundo Cardoso et al. (2007), um sistema de RA tem como características básicas o processamento em tempo real, combinação de objetos virtuais com o mundo real e o uso de elementos virtuais concebidos em 3D.

A principal diferença entre RV e RA é que a primeira cria um mundo virtual para que o usuário interaja e experimente novas sensações, já a segunda coloca objetos virtuais modelados em 3D (como por exemplo, gráficos 3D) no mundo real, a fim de complementá-lo. O uso de RV e RA como formas de visualização da informação possibilita maior interação com os dados.

3.3 Filtros bayesianos

Taveira (2008) afirma que o objetivo dos filtros bayesianos é automatizar o processo humano de reconhecimento de spam por meio de cálculos. Adeniji, Adigun e Adeyemo (2012) asseguram que, nesse método, os documentos são organizados em categorias e assim, é possível criar um índice de probabilidade spam para cada palavra presente no e-mail.

Segundo eles, os algoritmos de filtragem são divididos em duas partes, são elas a de treinamento e a de classificação. A de treinamento consiste em um conjunto de mensagens previamente rotuladas como spam ou não e nele é aplicado o algoritmo de aprendizado. A partir dos resultados obtidos com a fase anterior, são feitas as classificações das próximas mensagens.

3.4 Funcionamento do filtro e cálculos utilizados

Paul Graham (2004) propõe um filtro bastante simples e eficiente, atribuindo um *score* (pontuação) para cada palavra spam e logo após isso faz a combinação desses por meio de cálculos bayesianos.

Primeiramente, é separado um conjunto de spams e outro de não spams. Todas as mensagens de cada conjunto são varridas incluindo os cabeçalhos, HTML e Javascript embutidos. Caracteres alfanuméricos, traços, apóstrofes, cifrões são descartados e o restante é o chamado de token que é definido como cada palavra útil do e-mail.

É contado o número de vezes que cada token ocorre em cada conjunto. Feito isso, duas tabelas são geradas (uma para cada conjunto) que mapeiam o token e o respectivo número de ocorrências.

Depois é criada uma terceira tabela, mas dessa vez mapeando para cada token a probabilidade de um e-mail que o contém ser spam. A probabilidade é calculada de acordo com a Equação 1.

$$r_g = \min(1, \text{good}(w)/G) \quad r_b = \min(1, \text{bad}(w)/B)$$

$$P_{\text{spam}|w} = \max(0.01, \min(0.99, r_b/(r_g + r_b)))$$

Equação 1. Probabilidade de a palavra estar contida em um spam Graham (2004, p.123)

Na Equação 1, w é o token cuja probabilidade está sendo calculada, *good* e *bad* são as tabelas criadas no primeiro passo e G e B são os números de mensagens não spam e spam respectivamente.

Palavras que se encontram no conjunto de spams e não se encontram no conjunto de não spams, Graham (2004) atribui valor 0,99 a ela, caso contrário, atribui-se 0,1 e se ela não estiver contida em nenhum dos conjuntos o valor atribuído é 0,4.

Quando um novo e-mail chegar este é separado em tokens e em seguida é feita uma busca na base aprendida separando os vinte tokens mais interessantes (em que interessante é medido por quão distante as probabilidades estão distantes do neutro 0,5) são usados para calcular a probabilidade

de o e-mail ser spam. A probabilidade é calculada como mostra a Equação 2.

$$P_{spam} = \frac{\prod_{i=1}^{20} P_{spam|w_i}}{\prod_{i=1}^{20} P_{spam|w_i} + \prod_{i=1}^{20} (1 - P_{spam|w_i})}$$

Equação 2. Probabilidade de ser spam Graham (2004, p.124)

Na Equação 2, $P_{spam|w_i}$ é a probabilidade de a palavra pertencer a um e-mail spam.

O e-mail é classificado como spam, se o algoritmo acima resulta numa probabilidade maior que 0,9.

3.4 Metodologia utilizada

A metodologia utilizada para o desenvolvimento do trabalho é identificada como uma abordagem qualitativa, que tem como objetivo analisar e coletar dados de spams, bem como desenvolver um filtro que possibilite a classificação de e-mails como spam ou não e a visualização dos dados gerados pelo processo de classificação.

Depois de realizados esses estudos, foi definido o escopo do projeto e o início da implementação.

3.5 Recursos utilizados

Para o desenvolvimento da ferramenta foram utilizados os seguintes recursos:

- Banco de dados PostgreSQL 9.3, para a armazenagem dos dados coletados pela análise dos spams classificados;
- C# e JavaScript como linguagens de programação para o desenvolvimento da ferramenta;
- Microsoft Visual Studio 2013, ambiente de programação para desenvolvimento em C# e NetBeans 8.0 para o desenvolvimento em JavaScript;
- A biblioteca Three.JS para a montagem dos gráficos e as interatividades com estes;
- Renderizador Canvas e WebGL para a renderização dos gráficos propostos.

4 ESPECIFICAÇÃO DO PROJETO

Esta seção tem como objetivo apresentar o escopo, funcionalidades e diagramas relacionados ao projeto.

4.1 Escopo e funcionalidades

Este trabalho consiste na visualização em realidade virtual (RV) dos dados gerados por meio da filtragem de spams. Tal filtro foi implementado utilizando cálculos estatísticos bayesianos o que gerou uma base de dados com as informações

necessárias para a montagem das representações gráficas. O projeto foi dividido em duas partes: a primeira abrange o desenvolvimento do filtro bayesiano e a parametrização dos dados obtidos, já a segunda consiste na modelagem e exibição de gráficos em um ambiente RV.

Foi fornecida uma base de dados constituída de e-mails spams e não spams para a realização. Tal base pronta foi utilizada como entrada para o algoritmo de aprendizagem, o qual percorreu os e-mails fornecidos coletando palavras que caracterizam spam. Feito isso, foi atribuído um determinado valor para cada palavra encontrada, indicando a probabilidade desta ser um spam.

Uma vez concluído o treinamento do algoritmo de aprendizagem, o próximo passo é o carregamento dos e-mails que se pretende avaliar. Cada e-mail foi analisado por meio de cálculos estatísticos bayesianos visando classificar as palavras encontradas em seu conteúdo. Tais palavras foram confrontadas com a base aprendida visando atribuir o valor correspondente.

Ao mesmo tempo foram analisados outros parâmetros que possam identificar questões como origem e destino do e-mail, palavras mais comuns, porcentagens, dentre outros. Os dados parametrizados gerados pela análise foram armazenados em um banco de dados.

Depois de parametrizados tais dados, o usuário pode escolher quais informações deseja visualizar graficamente por meio de uma interface de configuração do sistema. Feito isso, será realizada a modelagem e a renderização dos gráficos de acordo com a especificação do usuário.

Os gráficos gerados podem ser escolhidos, dependendo do gráfico, em três opções: gráfico de barra, gráfico de área e gráfico de pizza. Ao passar o mouse em cima dos elementos do gráfico são exibidas as informações presentes em cada elemento. Além disso, é possível transladar o gráfico utilizando o botão direito do mouse, efetuar zoom com o scroll do mouse e rotacionar o gráfico no ângulo desejado mantendo o botão esquerdo do mouse pressionado.

Outra forma de interação com o gráfico é a possibilidade de verificar um valor de um dado componente do gráfico, para isso, basta, simplesmente, posicionar o mouse em cima do componente desejado.

4.2 Diagramas UML e Banco de Dados

O diagrama indicado na figura 6 apresenta o diagrama de classes do projeto e suas respectivas classes.

A classe *Aprendizagem* é responsável por realizar a aprendizagem a partir dos dois conjuntos de e-mails (spam e não spam) inseridos. Ela utiliza a classe *SeparaTokens* para percorrer os arquivos selecionados, separando-os em tokens. A classe *ClassificaEmail* atribui a porcentagem ao e-mail a partir do cálculo bayesiano e o classifica como spam

ou não. Juntamente com a classificação, a classe *AnalisaEmail* seleciona os campos interessados para o estudo e grava estes dados no banco por meio da classe *GravaBanco*.

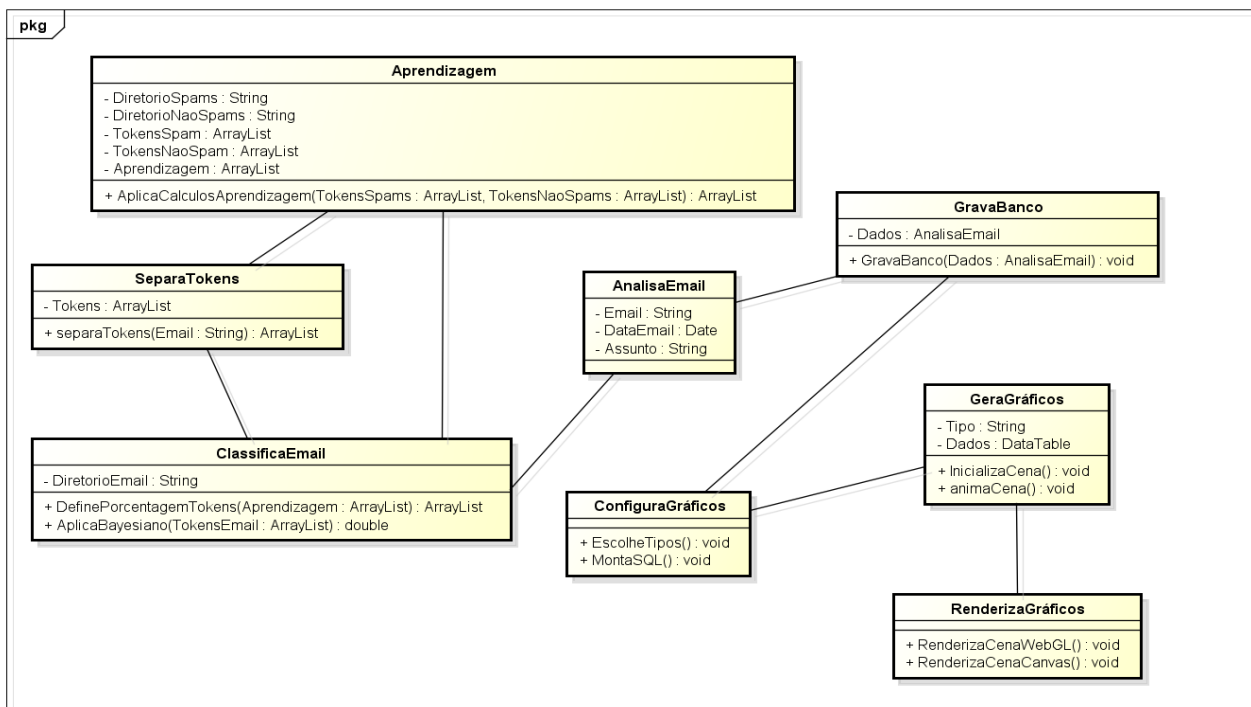


Figura 6. Diagrama de classes.

Por meio da classe *ConfiguraGráficos*, o gráfico escolhido pelo usuário é selecionado e o SQL é montado. A classe *GeraGráficos* coloca os objetos do gráfico na cena de acordo com o resultado dos SQLs e a classe *RenderizaGráficos* renderiza os objetos gerados na interface gráfica.

A Figura 7 ilustra o banco de dados com os dados dos e-mails que foram extraídos durante a classificação. Um e-mail é estruturado, basicamente, em cabeçalho, MIME e corpo da mensagem.

O cabeçalho contém informações como origem, destino, data, assunto, dentre outros. A tabela *eMail* corresponde à tabela do banco de dados em que as informações do cabeçalho serão guardadas.

A seção MIME possui informações técnicas da mensagem, como por exemplo, a prioridade do e-mail e se a mensagem possui algum anexo.

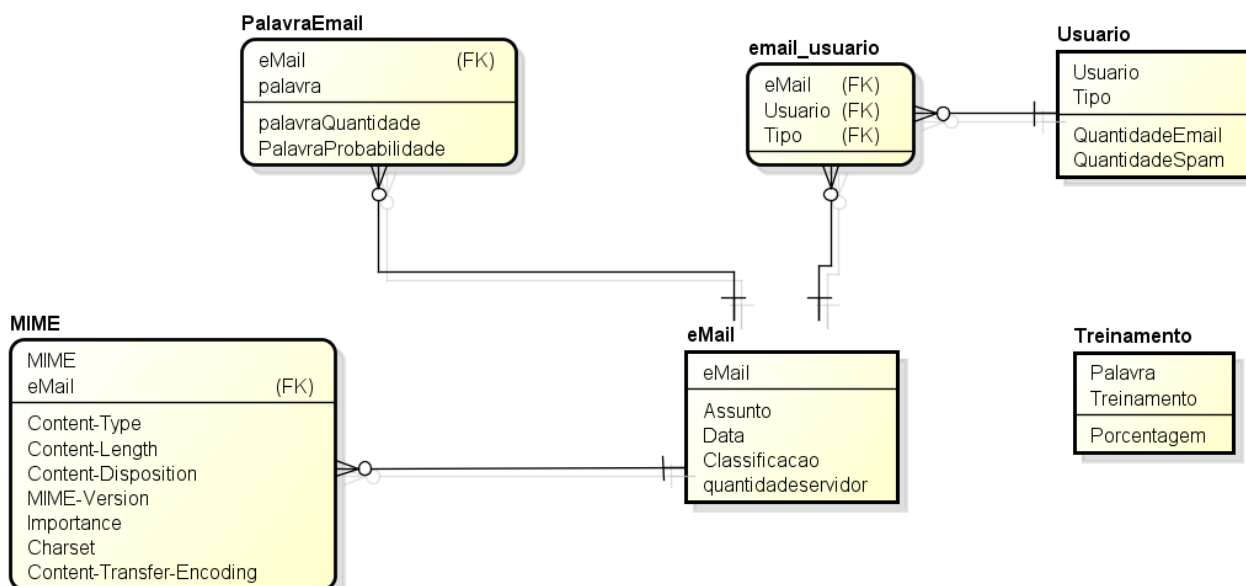


Figura 7. Relacionamento do banco de dados.

O MIME alterna de e-mail para e-mail, ou seja, uma informação pode estar presente no MIME de uma mensagem, mas não em outra. As informações gravadas no banco de dados são:

- **Content-Type:** especifica os tipos de conteúdo da mensagem. Por exemplo, se o Content-Type for text/plain, significa que a mensagem possui apenas textos simples; se for multipart/mixed, a mensagem possui textos simples e anexos.
- **Content-Lenght:** indica o tamanho da mensagem.
- **Contente-Transfer-Encoding:** indica a codificação da mensagem.
- **Content-Disposition:** possui informações como nome do arquivo (anexo), data de criação, tamanho do arquivo.

O corpo da mensagem é onde a mensagem propriamente dita se encontra.

A tabela *Palavra* no banco de dados contém as informações do corpo da mensagem e guarda as vinte palavras que foram utilizadas para a classificação e suas respectivas probabilidades de estar contida em um e-mail spam. O campo spam representa se a palavra é spam ou não.

Além das tabelas *eMail*, *MIME* e *palavra*, as demais tabelas que formam o banco de dados são apresentadas no esquema da figura 7. Um e-mail pode ter um ou mais usuários, isso é representado pela tabela *email_usuario*. A tabela *usuário* representa o emissor ou receptor do email e seus atributos são: usuário (e-mail), tipo (enviando ou recebendo), quantidade de e-mails

enviados ou recebidos e quantidade de spam enviado ou recebido. A tabela *palavra_usuario* guarda as palavras que determinado usuário enviou ou recebeu. Por fim, a tabela *treinamento* serve para facilitar o uso da ferramenta, para que o usuário não precise realizar a aprendizagem sempre que precisar utilizar a ferramenta.

O banco de dados guarda as informações necessárias para que a aplicação gere a visualização dos dados.

4.3 Arquitetura conceitual da aplicação

A figura 8 apresenta a arquitetura conceitual da implementação do protótipo.

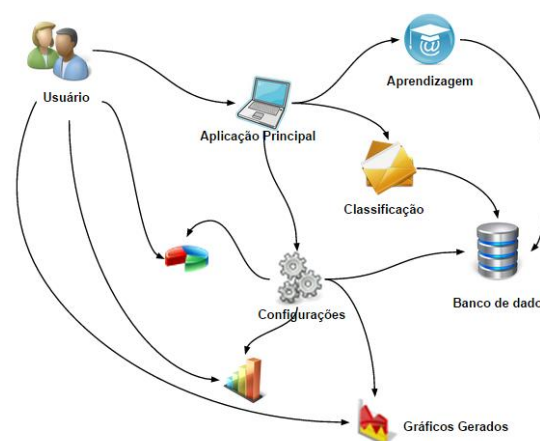


Figura 8. Arquitetura Conceitual.

A aplicação principal é por onde o usuário interage com o sistema desenvolvido, inserindo os conjuntos de e-mails para realizar a aprendizagem, classificação dos e-mails e a configuração dos dados realizados. A classificação e a aprendizagem armazenam os dados coletados

durante todos os processos. O banco de dados fornece informações para a configuração e montagem dos gráficos escolhidos pelo usuário e este pode interagir com os gráficos escolhidos da maneira desejada.

4.3.1 Aprendizagem

A figura 9 detalha o processo da aprendizagem descrito anteriormente na seção 3.4.

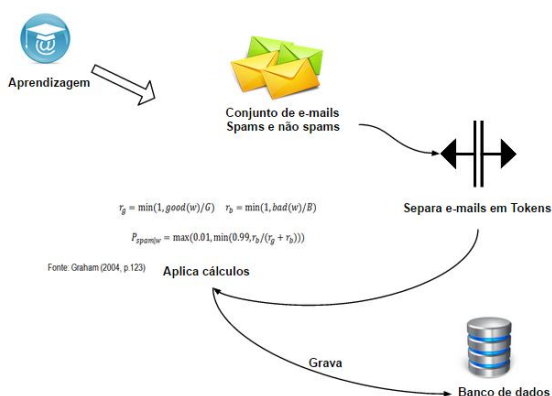


Figura 9. Banco de dados.

Os dados da aprendizagem são gravados no banco de dados. A tabela correspondente a esse processo é a *Treinamento*, apresentada na figura 7.

4.3.2 Classificação

Para realizar o processo de classificação, os e-mails são percorridos e separados em tokens, os quais são buscados na base de aprendizagem, visando identificar informações relevantes para a construção dos gráficos, os quais são gravadas no banco. A figura 10 detalha este processo da classificação de e-mails que já foi descrito na seção 3.4.

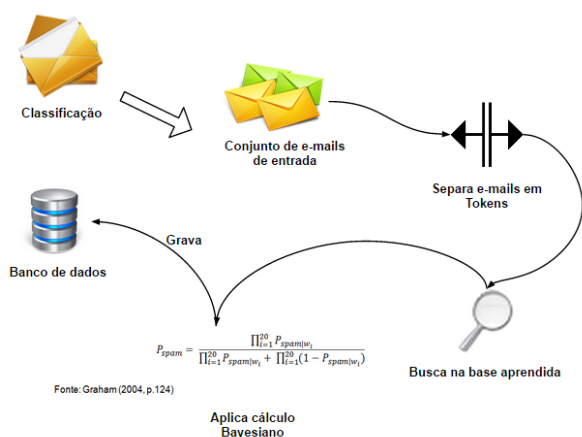


Figura 10. Classificação de e-mail

4.3.3 Visualização dos Gráficos

Uma vez preenchido o banco de dados, o usuário pode configurar o gráfico que deseja visualizar e escolher o tipo do mesmo. Feito isso, o banco de dados é consultado e em seguida é gerado um arquivo JavaScript com os dados buscados.

A partir dessas informações uma cena é criada e todos os objetos do gráfico posicionados nela. A cena é animada de acordo com as interações definidas. A cena é renderizada por canvas ou WebGL, dependendo do renderizador do navegador. A figura 11 detalha o processo da visualização dos dados gerados pela aplicação nos passos anteriores.

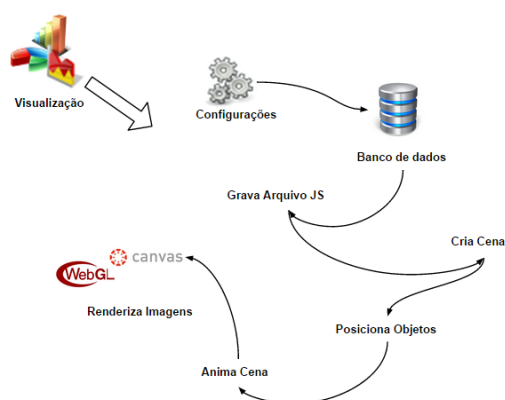







Figura 11. Visualização dos dados.

5 VISÃO GERAL DA FERRAMENTA

A interface principal da ferramenta disponibiliza na lateral direita, os ícones para as funcionalidades de aprendizagem , execução de testes , classificação dos e-mails  e para a visualização dos gráficos 3D interativos .

Para realizar a aprendizagem, deve-se clicar no ícone  e a aba aprendizagem será exibida, como mostrado na figura 12.

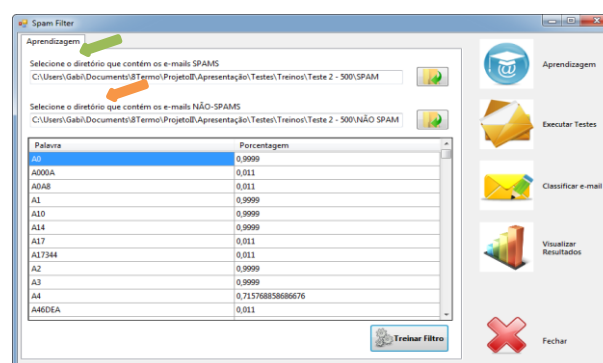


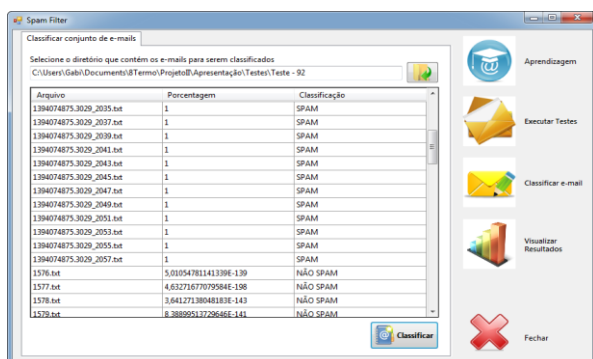


Figura 12. Aba para realização da aprendizagem.


No primeiro campo, na parte superior da tela (indicada pelas setas), deve ser selecionado diretório onde os e-mails spams se encontram e no segundo campo, o diretório onde se encontram os e-mails não spams. Para que a aprendizagem seja realizada, deve-se clicar no botão  Treinar Filtro na parte inferior da tela. Depois disso, será exibida uma tabela localizada na parte central da aba. Esta tabela contém as palavras da aprendizagem e suas respectivas probabilidades de estarem contidas em um e-mail spam.

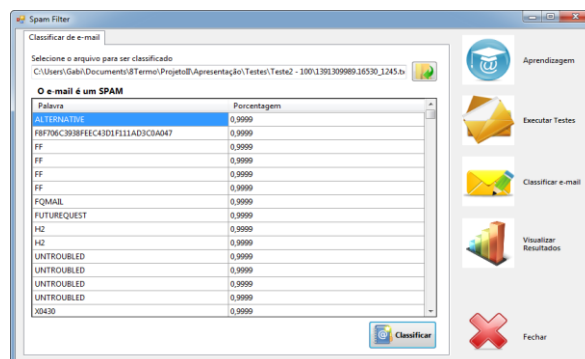
Somente após a execução da aprendizagem é possível classificar novos e-mails clicando no ícone  “Executar Testes”, esta função classifica todos os arquivos de um diretório. A figura 13 apresenta a tela de classificação de um conjunto de e-mails. Para realizar a classificação, selecione o diretório onde os e-mails se encontram e clique no botão “classificar”. A classificação será efetuada de acordo com a aprendizagem realizada anteriormente. Na tabela central da aba, o nome do arquivo do e-mail, a porcentagem do mesmo ser spam e a classificação deste e-mail, aparecerão respectivamente.



Arquivo	Porcentagem	Classificação
1394074875_3029_2035.txt	1	SPAM
1394074875_3029_2037.txt	1	SPAM
1394074875_3029_2039.txt	1	SPAM
1394074875_3029_2041.txt	1	SPAM
1394074875_3029_2043.txt	1	SPAM
1394074875_3029_2045.txt	1	SPAM
1394074875_3029_2047.txt	1	SPAM
1394074875_3029_2049.txt	1	SPAM
1394074875_3029_2051.txt	1	SPAM
1394074875_3029_2053.txt	1	SPAM
1394074875_3029_2055.txt	1	SPAM
1394074875_3029_2057.txt	1	SPAM
1576.txt	5,010547811413396-1,39	NÃO SPAM
1577.txt	4,6327677079584E-1,98	NÃO SPAM
1578.txt	3,64127138048183E-1,43	NÃO SPAM
1579.txt	8,3889951772646E-1,41	NÃO SPAM

Figura 13. Aba para a classificação de um conjunto de e-mails.

A figura 14 apresenta a tela de classificação de um único e-mail. Para realizar a classificação, clique no botão  Classificar E-mail. Selecione o arquivo que deseja classificar e clique no botão “classificar”. A classificação será efetuada de acordo com a aprendizagem, anteriormente, realizada.



Palavra	Porcentagem
A:TEPAA:TIVE	0,9999
FFP706C393BFEC43D3FL11AD3C0A47	0,9999
FF	0,9999
FF	0,9999
FF	0,9999
FF	0,9999
FQMAIL	0,9999
FUTUREQUEST	0,9999
H2	0,9999
H2	0,9999
UNTROUBLED	0,9999
UNTROUBLED	0,9999
UNTROUBLED	0,9999
UNTROUBLED	0,9999
XH3D	0,9999

Figura 14. Aba para a classificação de um único e-mail.



Para visualizar os dados gerados pela classificação das mensagens, deve-se clicar no botão  Visualizar Resultados e uma aba de configuração será aberta (figura 15) e nela será possível escolher os gráficos e seus tipos: de barra, de área e de pizza (indicadas pelas setas). Após esse passo basta clicar em  Gerar Gráfico.



Figura 15. Aba de configuração dos gráficos.

6 RESULTADOS E TESTES REALIZADOS

Foram realizados alguns testes a fim de validar o filtro implementado, testar a interatividade dos gráficos e analisar os dados armazenados durante a classificação em diferentes formas.

Para testar o filtro implementado, foram realizados testes com 200, 1000, 2000 e 7625 e-mails na aprendizagem e 92 e-mails no conjunto de teste. Na figura 16 pode-se visualizar o gráfico 3D de área com os resultados dos testes realizados. Quando a quantidade de e-mails na aprendizagem foi 200, 72 e-mails tiveram classificação correta, como mostra a figura 16-a (25 spams – figura 16-c e 47 não spams – figura 16-b), quando o número de mensagens foi alterado para 1000, 89 e-mails foram classificados corretamente (na figura 16 a, b, c e

d pode-se ver os valores mudando ao passar o mouse sobre o componente desejado e os valores serão apresentados).

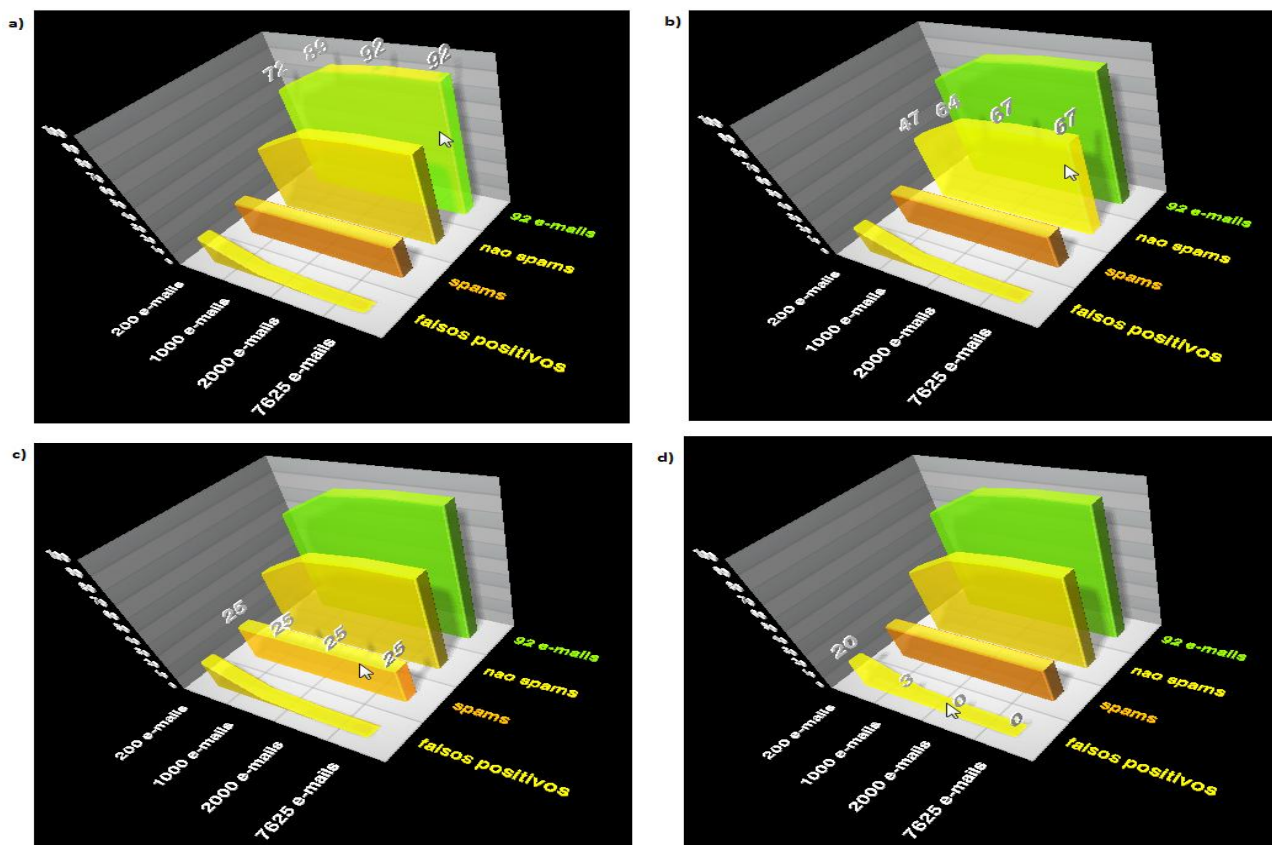


Figura 16. Testes realizados para a validação da ferramenta.

Depois disso, com as outras aprendizagens, não houve erros na classificação. É possível notar também a diminuição dos falsos positivos: no primeiro teste de 20 para 3 falsos positivos e nos dois seguintes zero (figura 16d).

Outro teste realizado tomou a aprendizagem de 7625 mensagens como base outro conjunto, dessa vez de 90 e-mails, foi classificado com o intuito de preencher o banco de dados. A figura 17 mostra a quantidade de e-mails spam e não spam classificados em forma de gráfico de pizza.

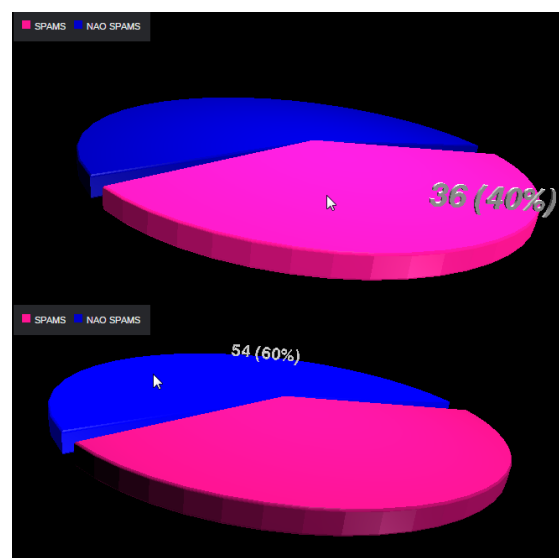


Figura 17. Quantidade de spams e não spams classificados.

Como se pode observar no gráfico, o conjunto é composto por 36 spams e 54 não spams, totalizando 40% e 60% do total, respectivamente.

Outros gráficos foram gerados como, por exemplo, a relação entre os diferentes charsets e a

classificação do e-mail como spam ou não. A figura 18 apresenta este gráfico.

Pode-se perceber, que os Charsets são característicos em spams e não spams. Por exemplo, o Charset “shift_jis” que é uma codificação de caracteres para o idioma japonês, aparece somente em spams, sendo assim, um gerente da rede analisada poderia bloquear todos os e-mails que tivessem o Charset “shift_jis” presente no MIME.

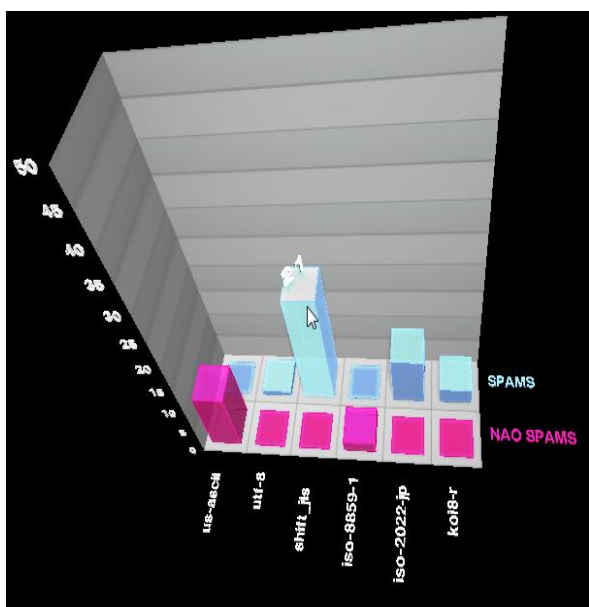


Figura 18. Relação entre os charsets e a classificação dos e-mails.

6.1 Comparação do modelo implementado com os trabalhos relacionados

A tabela 3 apresenta a comparação entre os trabalhos descritos na seção 2 do presente artigo. As colunas representam os pontos de comparação e as linhas cada trabalho comparado.

	Classificação de spam	Classificação Bayesiana	Exibição dos resultados	Uso de RA/RV	Interação com os gráficos	Configuração de Informações exibidas
2.1		X				
2.2	X					
2.3			X	X		X
2.4			X	X	X	X
Trabalho Implementado	X	X	X	X	X	X

Tabela 3. Comparação do modelo implementado com os trabalhos relacionados.

O trabalho 2.1 trata-se do Filtro de Conteúdo para Sistemas SMS Baseado em Classificador Bayesiano e Agrupamento por Palavras, o trabalho 2.2, Mecanismo Anti-Spam

Baseado em Autenticação e Reputação, o trabalho 2.3, Multidimensional Information Visualization Using Augmented Reality e o 2.4 Uso de realidade aumentada em ambientes virtuais de visualização de dados.

Nenhum trabalho relacionado possui tanto a classificação de e-mails, quanto a visualização da informação, sendo assim, eles estão divididos entre classificação spam e visualização da informação.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O filtro implementado possui um desempenho satisfatório em relação a classificação de spams. Entretanto, para obter melhores resultados é interessante combinar o filtro com outras formas de classificação, como por exemplo, listas brancas e negras. É importante ressaltar que e-mails spams mudam de destino para destino e de acordo com o tempo, sendo assim, é importante que a aprendizagem seja realizada de tempos em tempos e com conjuntos de e-mails relacionados ao destino onde o filtro será utilizado.

Em relação aos gráficos, eles permitem que os usuários interajam de maneira intuitiva com os resultados gerados pela classificação. A grande vantagem da utilização de gráficos 3D é a possibilidade de rotacionar, aplicar zoom, navegar entre gráficos, transladar um gráfico para a posição desejada e entre outras formas de iteração que asseguram uma análise de gráficos mais conscientes por parte de gerentes de rede, por exemplo.

Os gráficos gerados pela aplicação também auxiliam o estudo das mensagens spams, extraindo delas as principais características, o que proporciona melhoras aos filtros.

Com relação aos trabalhos futuros, há a necessidade de implementar mais gráficos, extraindo todas as informações do banco de dados e com formas diferentes de iteração, como por exemplo, possibilitar o usuário montar o próprio gráfico. Para isso, há a necessidade de estudo de mineração de dados.

Outra sugestão de trabalho futuro trata-se de implementar a ferramenta numa rede, possibilitando a análise em tempo real dos e-mails, assim como a conversão da ferramenta desktop para web, permitindo ao usuário fazer acesso remoto à ferramenta.

8. REFERÊNCIAS

ADENIJI, O. D.; ADIGUN, O.; ADEYEMO, O. O. An intelligent spam-scammer filter mechanism using bayesian techniques. **International Journal of Computer Science and Information Security**,

Nigeria, 3 mar. 2012. Disponível em: <<http://goo.gl/QpAaHP>>. Acesso em: 05 nov. 2013.

BELÉM, D. O. V. **Filtro de conteúdo para sistemas sms baseado em classificador bayesiano e agrupamento por palavras**. Belo Horizonte, 2009. Disponível em: <<http://goo.gl/AUExAt>>. Acesso em: 20 abr. 2014.

BUK, C. V. **Ambientes virtuais para visualização dinâmica de informação**. 2006. Dissertação (Ciência da Computação) - Universidade Metodista de Piracicaba, Piracicaba-SP.

CARDOSO, A. et al. **Tecnologias e ferramentas para o desenvolvimento de sistemas de realidade virtual e aumentada**. 2007. Disponível em: <<http://goo.gl/FPWZxt>>. Acesso em: 03 mai. 2014.

FREITAS, C. M. D. S. et.al. Introdução à visualização de informação. **Revista de Informática Teórica e Aplicada**, v.2, p.143-158, 2001.

GRAHAM, P. **Hackers & painters: big ideas from the computer age**. 1. ed. Sebastopol: O'Reilly Media, 2004. Cap. 8, p.121-129.

MEIGUINS, B. S. et al. Multidimensional information visualization using augmented reality. In: SYMPOSIUM ON VIRTUAL REALITY, 8. Belém-PA, 2006. Disponível em: <<http://goo.gl/ff64NV>>. Acesso em: 03 mai. 2014.

NASCIMENTO, H. A. D. Visualização de informação – uma abordagem prática. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25. JAI, 24. **Anais...** UNISINOS – São Leopoldo – RS, 2005.

TAVEIRA, D. M. **Mecanismo anti-spam baseado em autenticação e reputação**. Rio de Janeiro, mar. 2008. Disponível em: <<http://goo.gl/eu2wdF>>. Acesso em: 28 out. 2013.

ZORZAL, E. R.; CARDOSO, A.; KIRNER, C. **Estratégia para o desenvolvimento de aplicações adaptativas de visualização de informações com realidade aumentada**. Uberlândia, Nov., 2009. Disponível em: <<http://goo.gl/FC0uO1>>. Acesso em: 20 abr. 2014.