UNIVERSIDADE FEDERAL DE ITAJUBÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Diego Pires de Almeida

Um Sistema Antispam de Três Estágios

Dissertação submetida ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência e Tecnologia da Computação.

Área de Concentração: SISTEMAS DE COMPUTAÇÃO

Orientador: Benedito Isaías de Lima Lopes
Co-Orientador: Otávio Augusto Salgado Carpinteiro

Outubro de 2011
Itajubá-MG

Agradecimentos

Agradeço em primeiro lugar a Deus, por me conceder o dom da vida e a perseverança nos momentos difíceis.

Agradeço a meus país e a toda minha família pelo apoio incondicional frente a qualquer dificuldade e a minha noiva pela paciência, carinho e apoio concedidos durante meus momentos de ausência.

Agradeço também a todos os amigos e colegas que direta e indiretamente me apoiaram na superação de mais este desafio.

E agradeço em especial a meus orientadores, que tiveram papel fundamental na elaboração deste trabalho, me direcionando e apoiando nos momentos mais delicados.

A vocês, meu muito obrigado.

Resumo

Desde sua concepção, no final dos anos 80, a rede Internet vem consolidando-se como um dos mais eficientes meios para troca de informação. O correio eletrônico, ou email, tornou-se a principal ferramenta da Internet para troca de informações. Infelizmente, porém, o correio eletrônico tornou-se alvo de oportunistas, que se valem da praticidade e do baixo custo da ferramenta para disseminar conteúdo indesejado pela rede.

Emails spam ou spams são informações recebidas sem o consentimento prévio dos destinatários. Os spams, na maioria das vezes, possuem conteúdo publicitário, visando a promoção de serviços, produtos ou eventos. Acabam gerando problemas, tais como o desperdício de largura de banda da rede e perda de tempo e produtividade por parte dos servidores de emails e dos próprios usuários.

Este trabalho propõe um sistema antispam de três estágios. O primeiro, o pré-processamento, analisa o conteúdo do email em busca de padrões conhecidos e realiza eliminações e/ou substituições de conteúdo para simplificá-los e uniformizá-los. O segundo estágio, a seleção de características, determina as características mais relevantes do email, segundo duas classes de emails — Ham e Spam. O terceiro estágio, a classificação, classifica o email.

O sistema antispam é exaustivamente testado sobre três bases de dados públicas, disponíveis na Internet — SpamAssassin, LingSpam e Trec. O desempenho do sistema é avaliado segundo o percentual de classificações corretas nas duas classes — Ham e Spam. São avaliados também os tempos gastos no treinamento e teste do classificador neural, bem como os aspectos relacionados à manipulação dos emails presentes nas bases de dados.

Os resultados obtidos mostram-se bastante promissores. O sistema antispam apresenta ótimo desempenho nas três bases de dados empregadas.

Abstract

Since its conception, in the 80's, the Internet network has becoming one of the most efficient ways for information exchange. The electronic mail, or email, became the Internet main tool for information exchange. Unfortunately, however, the electronic mail became the target for opportunists, who take advantage of its practicality and low cost to disseminate unsolicited content over the network.

Spam emails or spams are information received without the previous consent of the users. Spams, in most cases, have advertising content, aiming at promoting services, products and events. They cause problems, such as the waste of network bandwidth and of time and productivity of network servers and users.

This work proposes an antispam system made up by three stages. The first, the pre-processing, analyses the email content searching for known patterns, and performs content eliminations and/or substitutions to simplify them and to make them uniform. The second stage, the feature selection, identifies the most relevant features of the email, according to two email classes — Ham and Spam. The third stage, the classification, classifies the email.

The antispam system is exhaustively tested on three public databases, available in the Internet — SpamAssassin, LingSpam and Trec. The system performance is assessed according to the percentage of correct classifications in both classes — Ham and Spam. It is also assessed the time spent in training and testing of the neural classifier, as well as the aspects related to the manipulations of the emails contained in the databases.

The results obtained are very promising. The antispam system has very good performance on the three databases employed.

Sumário

Lista de Tabelas

Lista de Figuras

1 Introdução

Desde o final dos anos 80, quando Tim Berners-Lee e sua equipe do CERN (*European Organization for Nuclear Research*, de Genebra) tiveram a idéia de desenvolver um sistema de hipertexto para facilitar a divulgação de pesquisas entre os cientistas, a finalidade da Internet e de todos os avanços relacionados a ela vem sendo criar o principal meio de comunicação do planeta.

Nesse sentido o correio eletrônico, ou email, ainda que criado antes mesmo da Internet, teve sua popularização com o crescimento da mesma, tornando-se uma das principais formas de intercomunicação pessoal existentes hoje. Segundo pesquisa realizada pela Meta Group (WEB, 2003), 80% dos usuários comerciais preferem o email ao telefone.

No entanto, como todo recurso de vinculação em massa que ganha destaque, logo surgem oportunidades para exploração indevida e, dado o baixo custo de se enviar um email, mensagens indesejadas podem ser disseminadas sem critério algum por toda a rede. São os chamados spams, mensagens eletrônicas que, na maioria das vezes, são de intuito publicitário, visando a promoção de serviços, produtos ou eventos, enviadas sem o consentimento prévio dos destinatários (JESSEN; CHAVES; HOEPERS, 2003). Portanto, no decorrer deste trabalho, as mensagens (emails), disseminadas com esta características serão denominadas spams, assim como as mensagens (emails), com conteúdo de interesse prévio do usuário serão denominadas hams.

1.1 A origem do termo Spam

Segundo, Templeton (Acessado em 22/01/2011), o primeiro spam pode ter surgido em maio de 1978, quando um funcionário da DEC, contratado para fazer propaganda de um novo produto, enviou para 320 endereços uma mensagem de divulgação.

A data oficial de nascimento do termo Spam é apontada, porém, como sendo 5 de março de 1994, quando dois advogados, Canter e Siegel, enviaram uma mensagem sobre

uma loteria de Green Cards americanos para um grupo de discussão sem foco no assunto. O ato causou revolta e indignação por parte dos seus usuários. Mas foi no dia 12 de abril de 1994 que o pior ocorreria. Utilizando um programa para envio de mensagens em massa, os advogados enviaram a mesma mensagem para vários grupos de discussão simultaneamente. As reações foram imediatas, acusando os responsáveis por violação das regras e princípios dos usuários da rede. O grande número de mensagens trocadas sobre o assunto comprometeu o desempenho da rede, causando um dos conhecidos efeitos colaterais do spam. O Apêndice A traz na integra a mensagem enviada.

A referência ao termo spam surgiu durante as inflamadas discussões sobre o ocorrido (ANTISPAM, Acessado em 15/02/2011). O fato relembrou uma cena do programa de TV do grupo inglês Monty Python, onde vikings inconvenientes estavam em uma lanchonete, repetindo diversas vezes a palavra "spam", referindo-se a um conhecido enlatado americano composto de presunto condimentado. A sensação de perturbação e incomodo foi utilizada como referência para a situação experimentada por usuários vítimas de spams. A figura 1 [1], ilustra a situação apresentada na propaganda.

Figura 1: Ilustração da propaganda vinculada ao SPAM®

O produto em questão, denominado SPAM® é fabricado pela Hormel Foods[2] desde 1930, que não aprova a associação de sua marca com algo tão nocivo à Internet e a seus

[1]Ilustração disponível em *http://www.antispam.br*
[2]Hormel Foods(http://www.hormel.com).

usuários. No site oficial do SPAM®[3], existe um texto esclarecendo que spam, grafado com letras minúsculas, diz respeito ao envio de mensagens não solicitadas, enquanto que "SPAM®", grafado com letras maiúsculas, refere-se marca registrada pela Hormel Foods.

1.2 O cenário nacional

O ano de 2009 foi marcado pelo significativo crescimento de spams originados em redes brasileiras. O gráfico da figura 2 mostra a evolução, entre os anos de 2003 a 2010, da quantidade de spams no país.

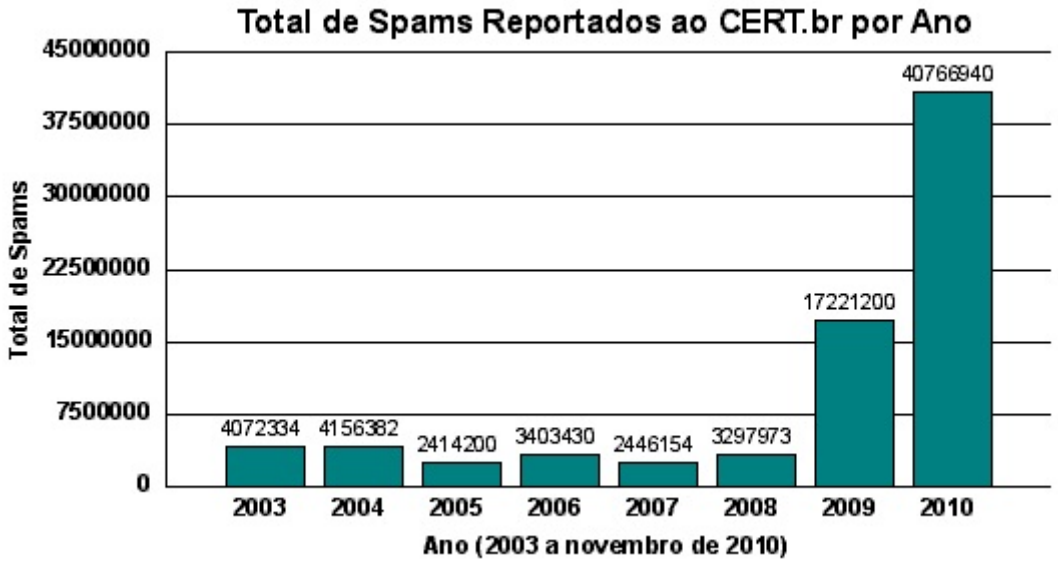


Figura 2: Evolução da quantidade de Spams no país

Segundo estudos realizados pelo Comitê Gestor da Internet no Brasil (CGI, Acessado em 15/02/2011) em 2009, o Brasil assumiu a liderança no número de máquinas comprometidas ou mal configuradas sendo abusadas por "spammers" do mundo todo. Fato que mostra que o problema não são os "spammers", mas sim o número crescente de máquinas de usuários finais, conectadas via banda larga, mas sem proteção, facilmente infectadas e abusadas por "spammers" de todo o mundo.

A partir disto, pode-se inferir que a solução para os abusos relacionados aos spams não está somente na busca de um conjunto de ações envolvendo operadoras de telecomunicações e políticas governamentais, mas sim também na conscientização dos usuários finais para adoção de técnicas para controle e bloqueio de mensagens indesejadas, contexto no

[3]Hormel Foods(http://www.spam.com).

qual este trabalho está inserido.

1.3 Principais problemas causados por Spams

Os spams, além do inerente desconforto de uma mensagem indesejada, ainda podem gerar prejuízos maiores para os usuários, como desperdício de largura de banda da rede e perda de tempo e produtividade por parte dos servidores, o que culmina em atrasos e, até mesmo, em não recebimento de mensagens legítimas (SILVA; MOITA; ALMEIDA, 2003). Os principais problemas relacionados à disseminação de spams podem ser resumidos em (ANTISPAM, Acessado em 15/02/2011):

Não recebimento de emails: Muitos provedores limitam o tamanho da caixa postal do usuário. Caso um grande número de spams seja recebido de forma não identificada, direcionado portanto para caixa de entrada do usuário, o mesmo corre o risco de ter sua caixa postal lotada após algum tempo sem gerenciá-la. Neste caso, passará a não receber emails até que possa liberar espaço em sua caixa postal. Um outro problema relacionado ao não recebimento de emails são os falsos positivos (detalhados na seção1.4.1 desse trabalho), quando o sistema antispam do usuário classifica erradamente uma mensagem legítima como spam.

Gasto desnecessário de tempo: Tempo dedicado pelo usuário para ler e identificar um email como spam, para somente então removê-lo de sua caixa postal.

Aumento de custos: Sempre quem paga o preço por um spam disseminado é quem o recebe. Desta forma usuários com cota de tempo ou de volume de transferência de dados são onerados cada vez que recebem um email spam, pelo qual não possuem interesse e, portanto, não estariam dispostos a pagar seu custo de recebimento.

Perda de produtividade: O recebimento de emails spam diminui a produtividade de todos que utilizam o email como ferramenta de trabalho, pois um tempo extra é dedicado a tarefa de recebimento e leitura dos emails, além do fato de que uma mensagem importante pode ser apagada por engano, não lida ou lida com atraso dado o volume de spams recebidos.

Conteúdo impróprio ou ofensivo: Como os spams são disseminados de forma aleatória, é bem provável que o usuário algumas vezes julgue impróprio e ofensivo o conteúdo da mensagem recebida.

Prejuízos financeiros causados por fraude: Spams estão sendo utilizados para

induzir usuários a acessar páginas clonadas de instituições financeiras ou a instalar programas maliciosos, projetados para furtar dados pessoais e financeiros, causando grandes prejuízos aos usuários.

1.4 Métodos para classificação de emails

Muitos métodos já foram empregados na tentativa de resolver o problema de emails indesejados, os famosos spams. Segundo Ozgur, Gungor e Gurgen (2004), porém, nenhum até agora obteve resultados satisfatórios, uma vez que, na tentativa de resolver o maior problema relacionado a identificação de spams, os falsos positivos, tais métodos propõem atenuantes que ao reduzirem os rigores de classificação, permitem a passagem de alguns spams, erroneamente classificados como hams.

Dentre os principais métodos empregados atualmente para a classificação de emails, podem-se citar as Listas Brancas, Listas Negras, Algoritmos Naive Bayesian, Support Vector Machines (SVMs), Boosting Trees, Aprendizado baseado em memória e as Redes Neurais Artificiais. Tais métodos podem ser classificados como estáticos ou dinâmicos, de acordo com a forma que interpretam os emails.

As seções a seguir trazem uma breve descrição do problema dos falsos positivos e dos métodos estáticos e dinâmicos para classificação de emails.

1.4.1 O problema dos falsos positivos

Os falsos positivos representam o principal problema relacionado à classificação de emails. Um falso positivo é um email legítimo, de interesse do usuário, classificado erroneamente como spam.

Este erro de classificação mostra-se severo na medida em que se analisa o comportamento da maior parte dos usuários de sistemas antispam. O que estes esperam da ferramenta é a garantia de filtrar apenas emails indesejados e, portanto, executam a ação de limpeza da caixa de quarentena sem verificação minuciosa de seu conteúdo. No entanto, se ali estiver presente um email legítimo com conteúdo aguardado pelo usuário, o mesmo será definitivamente eliminado, antes mesmo que o usuário tome conhecimento de sua existência.

Uma vez identificada a deficiência, o problema dos falsos positivos exige que o usuário verifique continuamente sua pasta de emails spam na tentativa de encontrar um email

legítimo, erroneamente classificado. O problema pode ser ainda pior quando o próprio servidor apaga os emails da pasta de spam, excluindo dessa forma um email ham sem que o usuário tenha tomado conhecimento dele.

A classificação de emails spam como ham também ocorre. Esta é conhecida como falso negativo, e também deve ser evitada. Os falsos negativos não apresentam, porém, implicações tão severas quanto as dos falsos positivos, já que o usuário pode identificar rapidamente a presença de um email spam em sua caixa de correio e movê-lo para a lixeira.

1.4.2 Métodos estáticos para a classificação de emails

Os métodos estáticos de classificação de emails geralmente contam com a intervenção do usuário no processo de classificação, uma vez que é este quem provê as informações para o servidor ou gerenciador de emails classificar uma mensagem como spam ou ham.

Estas informações são geralmente baseadas em listas, onde alguns endereços são classificados como fontes de spams (listas negras) ou quando somente emails provenientes de uma determinada lista são aceitos como hams (listas brancas). Tais informações podem ainda ser provenientes dos próprios emails. Em tal processo, o usuário seleciona algumas características presentes na mensagem pelas quais um email pode ser classificado como spam. Assim, as características são verificadas pelo servidor ou gerenciador e o email é automaticamente movido para a pasta de spams.

Uma outra forma de proteção empregada por alguns servidores de email é o reenvio de uma mensagem automática para qualquer endereço não presente na lista de contatos do usuário. Assim, o email só será validado como um ham quando a mensagem de retorno é identificada.

O grande problema desses métodos é que spammers do mundo todo estão em constante evolução, mudando o formato e o endereço de origem dos spams, exigindo do usuário um esforço contínuo na seleção de características e/ou endereços que tornem um email recebido válido ou inválido.

1.4.3 Métodos dinâmicos para a classificação de emails

A principal característica dos métodos dinâmicos é a de levar em consideração o conteúdo dos emails na classificação. A partir de uma base de dados pré montada de emails spams e hams, os classificadores analisam o "perfil" de cada uma das classes envolvidas e

passam a utilizar suas características como critério de classificação. A grande vantagem do emprego dessas técnicas é a independência em relação ao usuário e a possibilidade da evolução do classificador de acordo com a evolução dos spammers.

Como exemplo de métodos dinâmicos, podemos destacar os métodos baseados no Algoritmo Naive Bayesian, onde a principal idéia é calcular a probabilidade do email pertencer a uma classe (spam ou ham). O classificador Naive Bayesian apresenta um custo computacional bem inferior ao de outras técnicas baseadas em classificadores (ou filtros) bayesianos porque supõe que existe uma independência entre as palavras, ou seja, que a ocorrência de uma palavra em nada tem a ver com a probabilidade de ocorrência de uma outra. Isto torna o cálculo bem mais simples. Este método é empregado em alguns programas como o gerenciador de emails Mozilla Thunderbird.

Outros métodos dinâmicos também bastante utilizados são os baseados em: a) Support Vector Machines (SVMs), onde o objetivo mais geral é minimizar o erro na classificação, enquanto tenta-se maximizar a separação entre as classes; b) árvores de decisão que, com base em regras combinacionais, tentam classificar os emails a partir do conhecimento adquirido no processo de treinamento; c) algoritmos que se baseiam no armazenamento em memória (SAKKIS et al., 2003), onde o aprendizado acontece através das experiências armazenadas, em vez do conhecimento adquirido.

Além dos métodos dinâmicos citados, existem as Redes Neurais Aritificiais, descritas com maiores detalhes na seção 2.3.

1.5 Proposta de Trabalho

O método proposto neste trabalho para classificação de emails utiliza Redes Neurais Artificiais em conjunto com técnicas de filtragem e de seleção de características em emails, de forma a reduzir a complexidade do problema.

Desta forma, podemos definir o método proposto como sendo um sistema de três estágios. O primeiro analisa o conteúdo do email em busca de padrões conhecidos e realiza eliminações e/ou substituições de padrões para simplificação da classificação. O segundo utiliza técnicas de seleção de características, empregadas originalmente na área lingüística, para determinar as palavras de maior relevância para cada uma das classes analisadas (Spam e Ham). No terceiro e último estágio, de posse das informações geradas nos estágios anteriores, gera-se o vetor de entrada para a Rede Neural e a executa, para obter a classificação do email.

A técnica de "dividir para conquistar" mostra-se fundamental. Dividindo-se o problema em três estágios, reduz-se sua complexidade e aumenta-se o desempenho da rede neural em sua tarefa de classificação. Assim, os resultados obtidos superam os de outros trabalhos onde o problema é tratado em um único estágio.

Além de apresentar os resultados, o trabalho apresenta, igualmente, procedimentos não abordados por outros trabalhos na literatura, tais como:

- A medição do tempo de treinamento e teste do classificador neural;
- O emprego de bases de treinamento e teste com dimensões iguais, o que torna os resultados realísticos;
- A utilização de mais de uma Base de Emails, formadas a partir de fontes distintas e sem utilização de quaisquer tratamentos adicionais para os dados.

1.5.1 Conteúdo da Dissertação

A dissertação é dividida da seguinte forma. No capítulo 2, são apresentados a estrutura de um email bem como os conceitos fundamentais nos quais este trabalho se fundamenta. No capítulo 3, é apresentada uma revisão de trabalhos relacionados a área de sistemas antispam. O capítulo 4 apresenta a metodologia dos estudos realizados e os resultados obtidos. Por fim, o capítulo 5 conclui o trabalho apontando para direções de pesquisa futuras.

2 Revisão Teórica

Como mencionado anteriormente, o método proposto para classificação de emails consiste em um sistema de três estágios, sendo o primeiro deles o pré-processamento dos emails.

A grande importância do pré-processamento reside na simplificação computacional do problema, uma vez que tanto elimina características irrelevantes para a classificação, quanto agrupa as relevantes, uniformizando-as através de tags, reconhecidas posteriormente pelo estágio de classificação.

2.1 Formato dos emails e tags html

O formato de um email pode ser dividido em três partes:

- Envelope;
- Cabeçalho;
- Corpo.

Como em uma mensagem de correio, o envelope de um email contém as informações de remetente e destinatário. Estas informações são necessárias para que o servidor, ao receber uma mensagem, saiba a quem direcioná-la, da mesma forma que saiba a quem retornar uma mensagem de erro caso algum problema venha a ocorrer e a mesma não seja recebida pelo destinatário. O protocolo SMTP implementa essas características a partir dos comandos MAIL FROM e RCPT TO.

O cabeçalho de um email, da mesma forma que o envelope, tem como objetivo apresentar informações sobre a mensagem, elencando-as no formato nome: conteúdo em linhas consecutivas. Basicamente, o cabeçalho de um email registra as seguintes informações:

- **Return-Path:** representa o endereço para retorno da mensagem. Geralmente é copiado do envelope (MAIL FROM);
- **Received:** mostra a procedência do email, a data e a hora em que uma mensagem foi recebida. Através dos vários campos received presentes em uma mensagem podemos identificar o caminho percorrido pela mesma. No entanto, somente o endereço mais recente é confiável, uma vez que foi inserido pelo servidor.
- **From:** apresenta o remetente da mensagem e não é necessariamente igual ao apresentado no envelope ou no campo Return-Path.

Um ponto que deve ser destacado é que os campos From: e To: são nominais, o que significa que podem não refletir a real origem ou destino da mensagem. Como nem sempre todos os campos do cabeçalho são exibidos em um gerenciador de email, o usuário pode ser levado a crer que eles realmente refletem o remetente e destinatário da mensagem.

2.1.1 Corpo e conteúdo de um email

É no corpo do email onde a mensagem é transmitida, seja através de textos ou de anexos. A mensagem textual pode apresentar formato plano, sendo apresentada exatamente como foi transmitida, ou HTML, sendo processada pelo programa interpretador antes de ser exibida ao usuário.

Podemos encontrar ainda, no corpo de um email, extensões especiais chamadas MIME (LEE; HUI; FONG, 2002). A flexibilização permitida por estas extensões possibilita a inclusão de qualquer tipo de arquivo à mensagem, como textos em padrões alternativos ao ASCII e conteúdos multimídia.

Para tanto, um cabeçalho MIME no formato tipo:sub-tipo deve ser incluído antes de cada conteúdo a ser apresentado. Os ítens abaixo representam exemplos de cabeçalhos MIME:

- Content-type: audio/wav;
- Content-type: text/plain.

A tabela 1 mostra as combinações tipo:sub-tipo possíveis.

Tabela 1: Tipos/Subtipos MIME

Tipo	Subtipo	Descrição
Texto	Plano	Texto sem formatação
	Formatado	Texto com formatação inclusa
Imagem	Gif	Imagens em formato Gig
	Jpeg	imagens em formato Jpeg
Audio	Basic	Arquivo de som
Video	Mpeg	Vídeo em formato Mpeg
Aplicação	Octet-stream	Uma sequência de bytes não interpretada
	Postscript	Documento em Postscript
Mensagem	Rfc822	Mensagem MIME RFC 822
	Parcial	Mensagem foi dividida para transmissão
	Corpo Externo	A mensagem em si deve ser buscada na rede
Multipart	Mixed	Partes independentes apresentadas ordenadamente
	Alternative	Mesma mensagem apresentada em formatos diferentes
	Parallel	Partes visualizadas simultaneamente
	Digest	Cada parte é um nova mensagem RFC 822

A extensão MIME permite ainda uma combinação de mais de um formato (mais de um conteúdo MIME), denominado MIME multipart. Neste formato o conteúdo da mensagem é apresentado como uma árvore, processada recursivamente até a raiz.

A figura 3 exemplifica uma estrutura de árvore de conteúdo MIME Multipart.

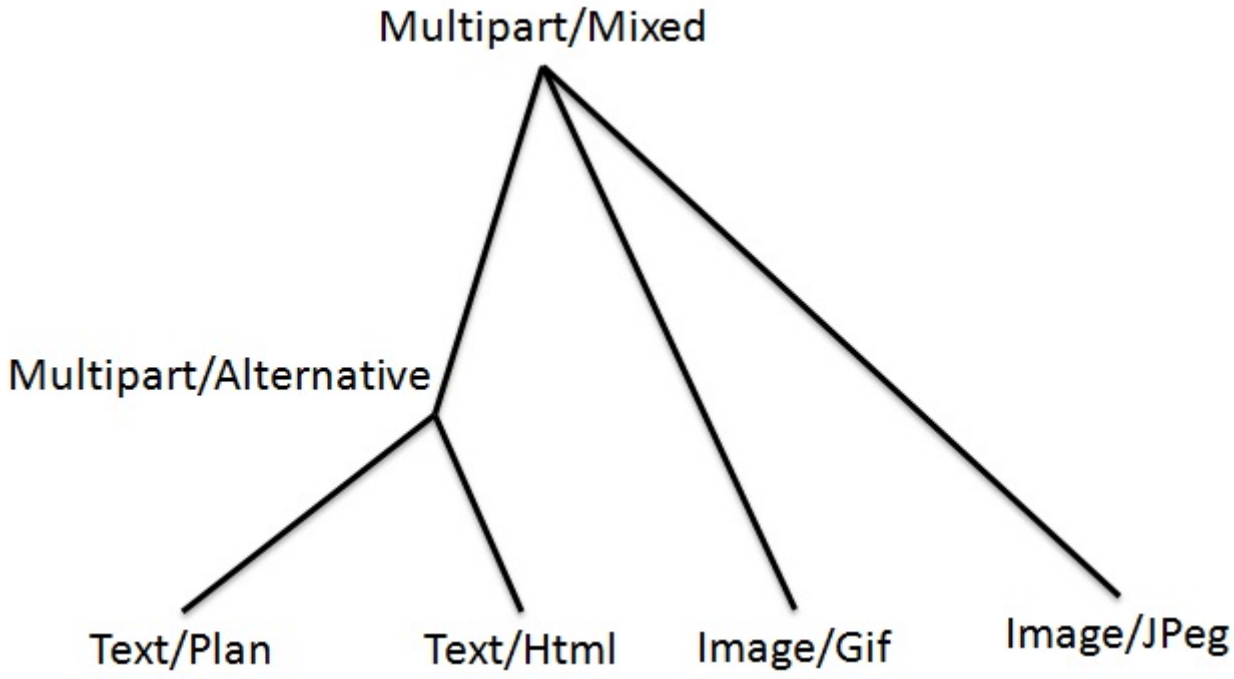


Figura 3: Exemplo estrutura MIME

2.1.2 Principais tipos de Spam

Uma vez definida a estrutura de uma mensagem, passamos a tratar da informação nela contida. Nesse sentido, cabe uma reflexão sobre os principais tipos de spam conhecidos hoje, bem como sobre as técnicas utilizadas por *spammers* em sua divulgação.

Segundo Antispam (Acessado em 15/02/2011) os principais tipos de spam identificados até hoje são:

- **Correntes** - o conteúdo de uma corrente geralmente traz uma história impactante, uma simpatia ou mesmo uma mensagem de sorte, solicitando ao usuário que repasse a mensagem um determinado número de vezes. Atualmente, spams do tipo corrente decresceram significativamente, entretanto ainda mostram-se freqüentes em grupos e listas de discussão. Alguns exemplos podem ser facilmente encontrados na internet;[1]

- **Boatos e Lendas urbanas** - são similares às correntes, porém diferem no conteúdo, geralmente trazendo histórias alarmantes e falsas alegando veracidade nos fatos, com depoimentos de conhecidos que passaram por situações semelhantes;

- **Propagandas** - O primeiro fator motivacional para a divulgação de spams, a propaganda comercial (UCE - Unsolicited Comercial Email), continua sendo hoje a principal razão de mensagens não solicitadas. A propaganda comercial aliada à capacidade publicitária proporcionada pela Internet pode, porém, acabar gerando o efeito inverso ao desejado. Empresas podem acabar comprometendo sua imagem e credibilidade por insistir em divulgar seu negócio por meio de mensagens não solicitadas;

- **Ameaças** - Caracteriza-se pelo fato de uma grande quantidade de mensagens geradas caluniando, difamando ou mesmo ameaçando um ou mais indivíduos. O simples fato da divulgação das informações em grande escala já caracteriza envio de spams e, quando a pessoa ou empresa envolvida sente-se lesada, cabe o registro de um Boletim de Ocorrência, eventualmente conduzindo a processos por calúnia e difamação;

- **Pornografia** - Os spams contendo conteúdo pornográfico figuram entre os mais comuns na rede. Dentre os prejuízos causados pelo recebimento de mensagens com conteúdo impróprio, podemos destacar o fato do interceptador do email poder ser menor de idade, e a divulgação de material com pedofilia. Neste último caso, a orientação é notificar imediatamente os órgãos responsáveis;

[1] http://www.quatrocantos.com/LENDAS/index_crono.htm

- **Códigos maliciosos** - são emails com conteúdo para convencer o usuário a executar um determinado código que, dentre outras consequências, pode corromper arquivos, infectar o computador com vírus, roubar senhas e informações pessoais;

- **Fraudes e golpes** - essas mensagens apresentam em seu conteúdo informações que levam os usuários a fornecer seus dados pessoais e financeiros, seja pela instalação de um software malicioso, seja pela reprodução de uma página falsa na Internet;

- **Spit e Spim** - o termo Spit refere-se ao "spam via *Internet Telephony*", com mensagens indesejadas propagadas via "telefones IP" (VoIP). O Spim, spam via *Instant Messenger*, é o termo empregado para as mensagens eletrônicas não solicitadas enviadas por meio de aplicativos de mensagens instantâneas, como *Microsoft Messenger* e o ICQ, por exemplo;

- **Spam via redes sociais** - com a popularização de redes sociais, como *Orkut* e *Facebook*, surge uma nova modalidade de spams, que se utiliza do objetivo desses aplicativos para disseminar pela rede boatos e propagandas.

2.1.3 O Pré-Processamento e o formato das Tags HTML

O pré-processamento consiste em um conjunto de aplicativos multi-plataforma desenvolvido em linguagem Java. Através do pré-processamento, a complexidade computacional do problema é reduzida, eliminando-se informações irrelevantes de forma a reduzir o conjunto de entradas para o classificador neural.

O processo de pré-processamento pode ser dividido em 2 estágios: limpeza dos emails e extração da lista de palavras. A lista de palavras é utilizada no segundo estágio pelos métodos de seleção de características para construção dos vetores de entrada para o classificador neural.

2.1.3.1 Limpeza dos emails

O primeiro estágio do pré-processamento é a limpeza do conjunto de emails da base de dados. Para tanto, o software desenvolvido em Java recebe como entrada, o caminho da base de emails e do diretório onde os arquivos "limpos" serão salvos.

Em seguida, o software inicia um minucioso processo de análise, onde cada email é processado individualmente. Primeiramente, o conteúdo do email é extraído e identificado como sendo simples ou multipart.

No caso do texto simples, podemos ter um conteúdo "puro" ou HTML. Quando um conteúdo de texto "puro" é identificado, o mesmo é separado em palavras (*tokens*) através dos seguintes caracteres delimitadores:

- Espaço;
- Nova linha;
- Tabulação;
- Exclamação;
- Interrogação;
- Vírgula;
- Ponto-e-vírgula.

Os *tokens* são então enviados à classe de identificação de padrões descrita na seção 2.1.5.

Quando um texto HTML é identificado, inicia-se o processo de identificação das tags. Uma tag HTML geralmente apresenta o seguinte padrão:

<nome_da_tag parametro1=valor1 parametro2=valor2> Texto da tag </nome_da_tag>

Além de tratamento e formatação multimídia, tags HTML podem conter complexas estruturas de categorização e disposição, incluindo até mesmo uma outra tag interna, exigindo, portanto, que a análise e o processamento destas estruturas ocorram recursivamente, em formato semelhante a uma árvore.

No pré-processamento proposto, as tags são divididas em três categorias de processamento distintas. A tabela 2 apresenta a divisão das tags entre as categorias.

O processamento das tags incluídas em cada uma das categorias apresentadas se dá da seguinte forma:

- **Categoria 1**: As tags presentes nesta categoria referem-se, em sua maioria, a estilos e formatação dos textos. A tag *tittle*, por exemplo, é utilizada por navegadores para apresentar o conteúdo na barra de navegação (título da página). Portanto, para fins de análise de um email, todo o conteúdo e parâmetros das tags da categoria 1 são descartados.

Tabela 2: Categorias de Processamento HTML

Tag	Categoria	Tag	Categoria
a	*3*	html	*2*
abbr	*2*	i	*2*
acronym	*2*	img	*3*
b	*2*	input	*3*
base	*3*	ins	*2*
body	*2*	label	*2*
br	*2*	li	*2*
button	*3*	map	*3*
caption	*2*	marquee	*1*
col	*2*	ol	*2*
comment tag	*1*	option	*2*
del	*2*	p	*2*
font	*3*	style	*1*
form	*3*	table	*2*
frame	*2*	textarea	*2*
h1-h6	*2*	title	*1*
head	*2*	tr	*2*
hr	*2*	var	*2*

- **Categoria 2**: As tags presentes na segunda categoria, por apresentarem conteúdos julgados parcialmente significativos para a correta identificação de um email, são processadas parcialmente, tendo apenas seus parâmetros descartados. Neste caso, a tag é substituída pelo conteúdo "!_in_nomeTag conteúdo". Um exemplo seria a tag *<body> Corpo do email </body>* substituída por "!_in_body Corpo do email".

- **Categoria 3**: As tags dessa categoria são processadas integralmente, o que implica que seus parâmetros e conteúdo são considerados integralmente para compor o email final resultante do processo de limpeza. Da mesma forma que apresentado na categoria anterior, a tag também é substituída, mas neste caso por !_in_nomeTag parâmetro conteúdo. Como exemplo temos a tag form *<form action=results.php> conteúdo </form>*, transformada em !_in_form action conteúdo.

Quando um conteúdo *multipart* é identificado, o primeiro passo é identificar se trata-se de um multipart/alternative ou de um *multipart* simples. Caso um conteúdo alternative seja identificado, somente a parte HTML será tratada.

O processamento segue com a identificação do conteúdo *multipart* sendo tratado. Neste ponto existem quatro possibilidades:

- A identificação de um texto simples, que leva ao tratamento apresentado anteriormente;
- A identificação de um conteúdo *alternative*, onde somente a parte HTML é tratada, também em acordo com o procedimento apresentado;
- A identificação de um conteúdo diferente dos conhecidos, onde uma tag informando o tipo do conteúdo é adicionada ao email;
- A identificação de um novo conteúdo *multipart*, que recursivamente deve ser identificado e tratado de acordo com o tratamento descrito.

2.1.4 Técnicas utilizadas por spammers

Com a evolução dos métodos para tratamento de emails maliciosos, avançaram também as técnicas utilizadas pelos *spammers* para o seu envio. A evolução das técnicas pode ser percebida tanto nos mecanismos de envio de spams quanto no tratamento dos conteúdos enviados.

2.1.4.1 Evolução nas técnicas de envio de mensagens

Com a evolução dos métodos estáticos para detecção de spams, como listas brancas e negras, os *spammers* foram obrigados a encontrar formas alternativas para enviar suas mensagens. Em Antispam (Acessado em 15/02/2011) podem ser vistas as principais evoluções nas técnicas utilizadas pelos *spammers*, tais como:

- **Programas de envio de email em massa**: também conhecidos como *bulk mailing* ou *mass mailing*, estes programas são fáceis de obter e podem ser utilizados para enviar emails através de máquinas mal configuradas, onde são instalados softwares maliciosos de manipulação. Endereços de máquinas com essas características são comercializados na Web, muitas vezes pelos próprios criadores dos serviços de envio de emails em massa;
- **Spam zombies**: computadores comprometidos por códigos maliciosos que, uma vez instalados, permitem que *spammers* utilizem a máquina para o envio de spam, sem o conhecimento do usuário;
- **Vírus propagados por email**: são normalmente executados como anexos de emails maliciosos. Um vez instalados, esses programas infectam o computador hos-

pedeiro, enviando cópias de si mesmo para os contatos encontrados no computador do usuário;

- **Abuso de formulários e scripts na Web**: consiste em spams enviados através de serviços na Web de transmissão do conteúdo de formulários por email. Segundo Antispam (Acessado em 15/02/2011), spams enviados a partir de servidores Web mal-configurados são dificilmente contidos pelas práticas atuais de contenção de spam. É muito importante, portanto, que administradores de serviços Web configurem corretamente seus servidores, evitando que sejam abusados por *spammers*;

- **Uso de sites comprometidos**: alguns *spammers* utilizam servidores comprometidos para enviar spams.

2.1.4.2 Evolução nas técnicas de tratamento dos conteúdos

Da mesma forma que ocorrido com as técnicas de envio de spams, os conteúdos dos emails maliciosos também tiveram que evoluir para tentar "enganar" os sistemas antispam. Segundo Cournane e Hunt (2004), algumas das principais técnicas utilizadas por *spammers* nesse sentido são:

- **Esconder palavras com caracteres inválidos** - consiste no fato de utilizar caracteres especiais, como pontos, traços e espaços, no meio de palavras, como *Viagra* e *Money*, sem prejudicar seu entendimento por parte do usuário. Como exemplo, a palavra "v.ia.g.ra";

- **Uso de tags HTML inválidas** - consiste no uso de tags HTML inexistentes, com o texto que o *spammer* deseja exibir. Exemplo, <Isso não é um tag HTML válida>

- **Uso de texto invisível** - consiste em usar técnicas de formatação HTML para esconder um texto legítimo, que apesar de não ser mostrado ao usuário, será processado e confundirá a classificação do sistema antispam. Como exemplo, temos a seguinte tag que mostra o texto ham na mesma cor de fundo da tela, não possibilitando, portanto, sua visualização por parte do usuário: *texto não visível para o usuário.;*

- **Uso de tag HTML em branco** - essa técnica é utilizada pelo *spammer* quando o verdadeiro conteúdo da mensagem encontra-se em um complemento, ou seja, o conteúdo HTML é apresentado sem texto plano, reservando-se ao complemento a

apresentação do significado. Um exemplo seria um texto apresentado dentro de uma imagem acompanhado de uma tag HTML em branco;

- **Uso de comentários HTML** - consiste em utilizar a tag de comentário HTML para esconder o verdadeiro significado de uma palavra. Isto faz com que o gerenciador, ao processar o texto, o apresente de forma perfeita para a compreensão da mensagem. Um exemplo pode ser expressado pela seguinte tag, onde a palavra *Money* é camuflada: *<HTML><BODY>mon<!-comentário não visível na tela ->ney</BODY></HTML>;*

- **Uso de texto redundante** - essa técnica apresenta um conteúdo multipart alternative, porém com um texto spam em formato HTML e um texto ham em formato plano. Isto confunde os sistemas antispam, e o conteúdo spam acaba sendo apresentado ao usuário;

- **Uso de acentuação incorreta** - consiste em usar acentuação incorreta para tentar enganar os sistemas que não retiram a acentuação para análise dos emails.

2.1.5 Detecção de Padrões Conhecidos

O processo de detecção de padrões tem início logo após a fase de detecção do conteúdo de email (texto plano, HTML, MIME, etc), quando o texto é separado em tokens para análise.

O primeiro tratamento realizado em cada *token* é a remoção de qualquer acentuação e a substituição dos caracteres em maiúsculo por caracteres em minúsculo. Neste estágio também é identificado se o token sendo tratado pertence ao cabeçalho do email ou não.

Tem início então a fase de identificação nos *tokens* de alguns padrões conhecidos utilizados por *spammers*.

- **Padrão XXX igual YYYY** - Ocorre quando temos atribuições que não agregam novo sentido ou informação ao conteúdo sendo analisado. Exemplo: *<table color=blue>* a saída seria !_in_table color e qualquer outra tag do tipo *<table color=X >* seria substituída por !_in_table color;

- **Endereços de emails** - todos os endereços de email encontrados durante a análise são substituídos por "!_EMAIL";

- **URLs** - da mesma forma que padronizado para emails, todo *hyperlink* (URL) encontrado durante a análise é substituído por "!_LINK";

- **Caracteres inválidos em meio a palavras** - toda vez que um padrão do tipo letra, caractere, letra, caractere,... é encontrado, o mesmo é substituído por "!_HIDEWORDS". Exemplo v-i-a-g-r-a, é substituído por "!_HIDEWORDS";

- **Espaço adicionado em meio a palavras** - da mesma forma que ocorrido com os caracteres inseridos em meio a palavras, quando um padrão do tipo letra, espaço, letra, espaço, letra, espaço, letra, espaço, ... é encontrado o conjunto é substituído por "!_HIDEWORDS". Exemplo: "v i a g r a" é substituído por "!_HIDEWORDS";

- **Palavras muito grandes** - quando uma seqüencia de tamanho anormal (normalmente sem sentido) é identificada pelo software, é substituída pelo identificador "!_BIGTEXT";

- **Assunto suspeito** - quando é encontrado no campo SUBJECT do cabeçalho uma sequência de números e caracteres, com intuito de burlar o tratamento de alguns sistemas, os mesmos são identificados e substituídos por "!_NUMERO_SUBJECT";

- **Dinheiro e Porcentagem** - qualquer referência encontrada sobre quantidades monetárias é substituída por "!_MONEY", da mesma forma que qualquer referência a percentuais encontrados é substituída por "!_PORCENTAGEM".

2.2 Métodos de Seleção de Características

Após o processamento e limpeza dos emails, ou seja, após o primeiro estágio do sistema antispam, tem início o segundo estágio, que consiste em escolher as características (palavras e tags inseridas) mais representativas dos emails para compor o vetor de entrada do classificador neural (detalhado na seção 2.3).

A escolha das características mais representativas é feita usando-se Métodos de Seleção de Características, amplamente utilizados na classificação de textos (YANG; PEDERSEN, 1997), o objetivo destes métodos é ponderar a relevância das características dos emails, sejam elas palavras, imagens, tags ou atributos Html, mais relevantes para cada uma das classes envolvidas na classificação.

A grande contribuição do estágio de seleção de características é diminuir a complexidade de análise do problema, limitando-a aos fatores mais relevantes e, como consequência,

reduzindo a dimensão do vetor de entrada do classificador neural. Sem este estágio, os vetores de entrada teriam dimensões inviáveis para processamento, ou pior, as características seriam determinadas sem uma ponderação formal, o que poderia resultar em perda de características importantes para a classificação, como mostrado na seção de análise dos resultados.

Este trabalho considera três métodos de seleção de características, escolhidos por sua popularidade e por terem demonstrado bom desempenho em outros trabalhos, discutidos no capítulo 3. Os métodos considerados foram:

- *Frequency Distribution* (DF);
- *Chi-Quadrado Statistic*;
- *Mutual Information* (MI)

2.2.1 Frequency Distribution (DF)

A distribuição de freqüência (DF) mede o grau de ocorrência de um elemento w em um conjunto C. Se w é uma característica, a distribuição de frequência da característica w é 2.1:

$$DF(\omega) = \frac{N \quad [\omega \in \{spam, ham\}]}{T} \tag{2.1}$$

O numerador representa o número de ocorrências de uma determinada característica nos conjuntos de spam e ham, e T representa o número total de características existentes nesses conjuntos. Após a classificação, são escolhidas as características com maior número de ocorrências, ou seja, com o maior índice DF.

É importante ressaltar que em nenhum momento o método DF pondera a relevância de uma característica em um determinado conjunto, somente a quantidade de vezes em que ela apareceu. Por este motivo, em alguns casos, a informação contida em características de baixa ocorrência, mas que são significativas para determinação da classe a qual o email pertence podem ser deixadas de lado.

Em resumo, o método DF se mostra bastante eficaz para a redução da complexidade do problema e, ainda que não considere a importância de uma característica para um conjunto, apresenta resultados extremamente positivos considerando-se sua simplicidade computacional.

2.2.2 Chi-Quadrado Statistic

A distribuição *chi-quadrado* mede o grau de dependência entre um elemento e e um conjunto S (PAPOULIS; PILLAI, 2001). Se w é uma característica e C um conjunto de duas classes – spam e ham –, a distribuição *chi-quadrado* da característica w é dada por 2.2:

$$\chi^2 = P(spam) \cdot \chi^2(\omega, spam) + P(ham) \cdot \chi^2(\omega, ham) \tag{2.2}$$

onde $P(spam)$ e $P(ham)$ são as probabilidades de ocorrência de emails spam e ham, respectivamente. A distribuição *chi-quadrado* para a característica w e classe c é dada por 2.3:

$$\chi^2(\omega, c) = \frac{N \cdot (kn - ml)^2}{(k+m) \cdot (l+n) \cdot (k+l) \cdot (m+n)} \tag{2.3}$$

Onde:

- N é o número total de emails no conjunto c;
- k é o número de emails pertencentes ao conjunto c que contém a característica w;
- n é o número de emails pertencentes ao conjunto c' que não contém a característica w;
- m é o número de emails pertencentes ao conjunto c que não contém a característica w;
- l é o número de emails pertencentes ao conjunto c' que contém a característica w;
- Sendo denominado c' o conjunto complementar ao conjunto c.

As características são então ordenadas pelos valores *chi-quadrado* apresentados. As características com maior índice são consideradas de maior relevância, e portanto selecionadas para a análise.

2.2.3 Mutual Information (MI)

Da mesma forma que o método *chi-quadrado*, o método MI também considera a relevância de cada característica para um determinado conjunto. A equação que define o método é mostrada pela equação 2.4:

$$MI(\omega) = \sum_{f=\{\omega,\varpi\}} \sum_{c=\{spam,ham\}} P(f,c) \log_2 \frac{P(f,c)}{P(f) \cdot P(c)} \tag{2.4}$$

Onde $P(f,c) = P(c) * P(f|c)$ é a probabilidade de f e c ocorrerem simultaneamente. As características são então ordenadas com base no cálculo de MI. São consideradas mais relevantes para análise, as características com valores mais altos.

2.3 Redes Neurais Artificiais

A busca por uma maneira de replicar o neurônio humano de forma artificial tem sido o objetivo de muitas propostas. Sua origem data do ano de 1943, quando o neurofisiologista Warren McCulloch, do MIT, e o matemático Walter Pitts da Universidade de Helenos, publicaram um trabalho sobre "neurônios formais", onde era feita uma analogia entre o funcionamento de células nervosas vivas e o processo eletrônico.

Tamanha euforia e investigação acabaram criando um novo segmento na área de Inteligência Artificial, as Redes Neurais Artificiais, capazes de reconhecer informações e produzir respostas seguindo a teoria de aprendizagem cognitiva, se adaptando a mudanças de controle, classificação e processamento ao longo do tempo (MALCON; XEREZ, 1996).

As redes neurais artificiais tentam modelar o comportamento do cérebro humano, através de neurônios artificiais, lineares ou não, que simulam o comportamento dos neurônios humanos. Mais precisamente, as redes neurais artificiais propõem um modelo matemático baseado em nós, conexões e pesos, que modela as conexões sinápticas nervosas entre os neurônios humanos.

Segundo Haykin (1999), as Redes Neurais Artificiais são semelhantes ao cérebro humano em dois aspectos:

- A experiência é a fonte do conhecimento adquirido;
- O conhecimento adquirido é armazenado nas sinapses.

O componente fundamental do Sistema Nervoso humano é o Neurônio que, apesar de ser individualmente responsável por processos extremamente simples, tem em suas complexas estruturas de conexões e ramificações todo potencial para realização das tarefas do Sistema Nervoso humano.

A figura 4 ilustra um neurônio biológico. Podemos perceber que o mesmo é dividido em três regiões distintas: corpo, dentritos e axônio.

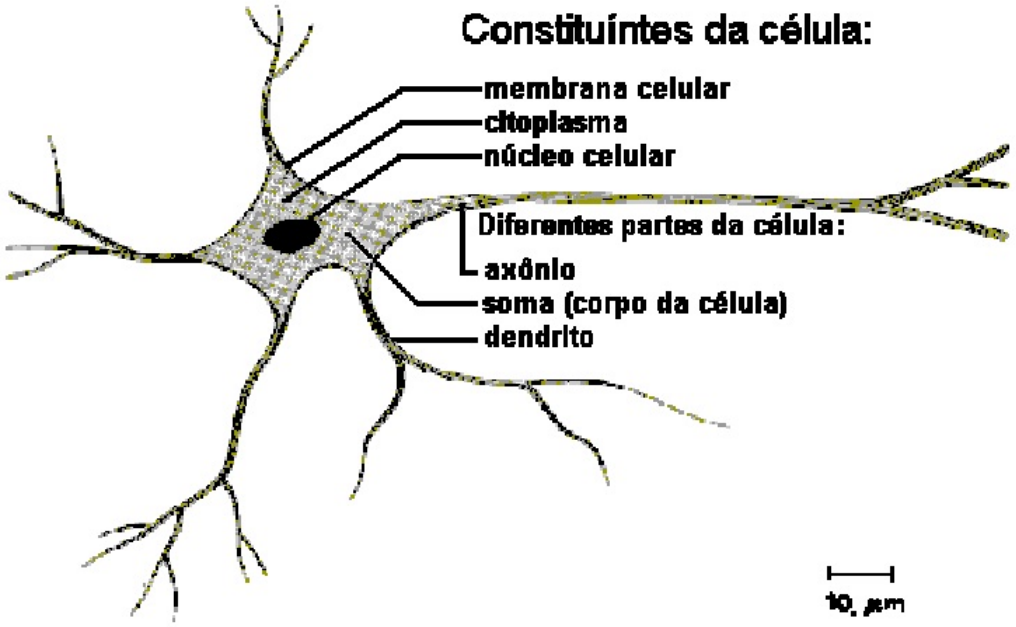


Figura 4: Neurônio Biológico

O corpo celular, chamado Soma, é onde se encontra o núcleo do Neurônio. O axônio, uma fibra nervosa de superfície lisa com poucas ramificações e maior comprimento, é responsável pela transmissão da informação para outros neurônios. Os dendritos, que têm aparência de árvores e possuem superfície irregular e muitas ramificações, atuam como receptores nesta comunicação.

A comunicação, ou interação, é chamada sinapse, e é caracterizada por um processo químico no qual são liberadas substâncias transmissoras que se difundem pela junção sináptica entre neurônios, o que causa aumento ou queda no potencial elétrico do neurônio receptor. Resumindo, uma sinapse é a conexão entre neurônios o que implica em excitação ou inibição do neurônio receptor (HAYKIN, 1999).

Da mesma forma que seu correspondente biológico, o principal componente das Redes Neurais Artificiais é o Neurônio Artificial, cujo modelo pode ser dividido em duas regiões distintas:

- Região de conexão entre as sinapses, que pode ter tanto valores positivos quanto negativos;

- Região de saída do sinal do neurônio, a qual varia em amplitude com relação à somatória dos sinais de entrada.

A figura 5 ilustra um neurônio artificial, com valores de entrada $[x_1, x_2, ...x_i]$, conjunto de pesos $[w_1, w_2, ...w_i]$, função de ativação f e saída y.

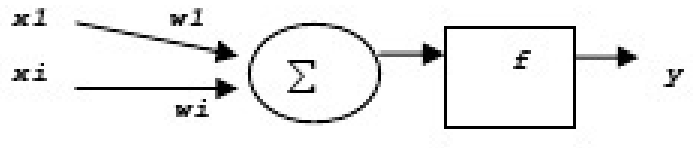


Figura 5: Neurônio Artificial

O neurônio artificial pode ser descrito pela equação 2.5 (HAYKIN, 1999).

$$y_k = \rho(\sum_{i=1}^{n} x_i w_{ki}) \tag{2.5}$$

Onde:

- y é a saída do neurônio;
- ρ é a função de ativação;
- $x_1, x_2, ..., x_n$ são os sinais de entrada do neurônio;
- $w_{k1}, w_{k2}, ..., w_{kn}$ são os pesos sinápticos do neurônio em questão (neurônio k).

Portanto, o neurônio artificial imita o funcionamento do neurônio biológico por meio de entradas e pesos, que representam os estímulos e pesos das conexões sinápticas, e pela função de ativação que simula o processo químico que libera substâncias que excitarão ou inibirão os próximos neurônios.

2.3.1 Processamento nos Neurônios

Cada Neurônio Artificial é responsável por realizar um processamento simples, que se inicia com uma entrada (ou estímulo) e termina com o processamento de um novo nível de ativação (RUSSEL; NORVIG, 1995). O processamento basicamente ocorre em duas etapas.

Na primeira, o estímulo é multiplicado pelo peso da conexão sináptica. Os resultados de cada estímulo são então somados, determinando a entrada para a função de ativação do neurônio. Na segunda etapa, a função de ativação é aplicada ao somatório da entrada, obtendo-se a saída do neurônio, representada na equação 2.6.

$$y = f(\sum x_i w_{ji}) \tag{2.6}$$

A linearidade, ou não, do neurônio é determinada pela função de ativação. Algumas das principais funções de ativação utilizadas são:

- Função Linear, equação 2.7, figura 6.

$$f(x) = ax \tag{2.7}$$

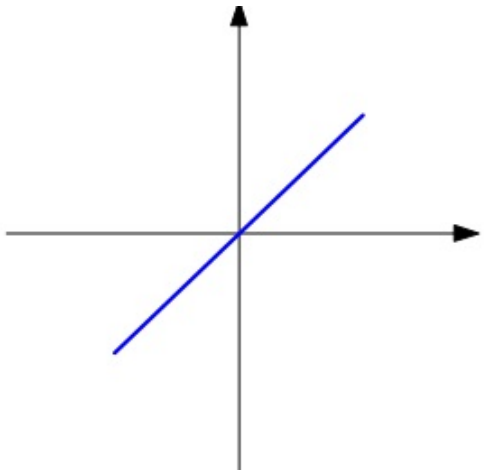


Figura 6: Função Linear

- Função Degrau (utilizada para valores binários de saída), equação 2.8, figura 7.

$$f(x) = 1 \quad se \quad x > 0, \quad ou \quad 0 \quad se \quad x \leq 0. \tag{2.8}$$

- Função sigmóide (é contínua, diferenciável, e representa transição gradual entre dois estados), equação 2.9, figura 8.

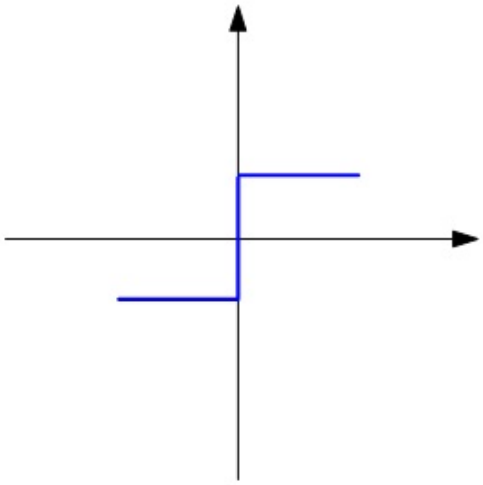


Figura 7: Função Degrau

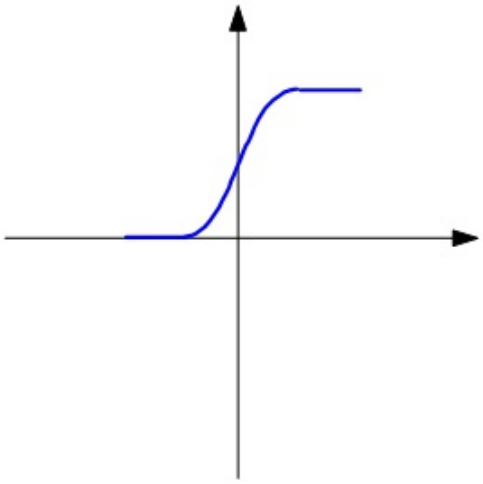


Figura 8: Função Sigmóide

$$f(x) = \frac{1}{1 + \exp^{-x}} \tag{2.9}$$

- Função tangente hiperbólica (função sigmóide que contempla também valores negativos), equação 2.10, figura 9.

$$f(x) = \frac{1 - \exp^{-x}}{1 + \exp^{-x}} \tag{2.10}$$

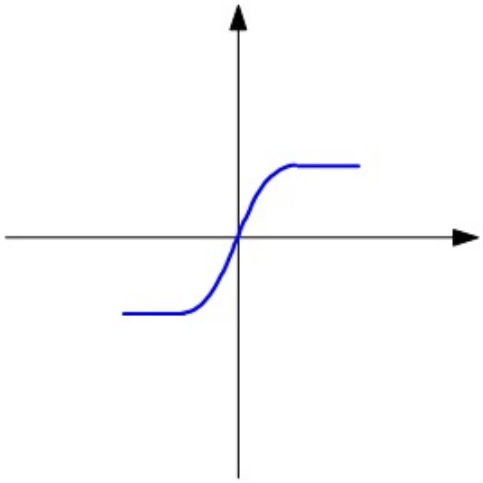


Figura 9: Função Tangente Hiperbólica

2.3.2 Redes MLP e Backpropagation

As primeiras Redes Neurais desenvolvidas possuíam seus neurônios dispostos em uma única camada. Tais modelos ficaram conhecidos como Perceptrons (figura 10). Porém, verificou-se que o modelo Perceptron falhava ao tentar classificar problemas não-linearmente separáveis.

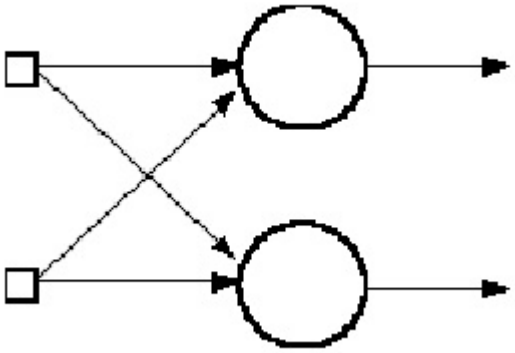


Figura 10: Modelo Perceptron

Com o intuito de resolver problemas mais complexos (incluindo os não-linearmente separáveis), surgiram as redes MLP (*multi-layer-percetron*). As MLPs possuem uma (ou mais) camada(s) intermediária(s) (camadas *hidden*). A figura 11 ilustra uma rede MLP, onde são destacadas três camadas funcionalmente distintas:

- Camada de entrada (CE), onde os padrões são apresentados à rede;

- Camadas intermediárias (CI), onde a maior parte do processamento é realizado;
- Camada de saída (CS), onde as respostas dos padrões são apresentadas.

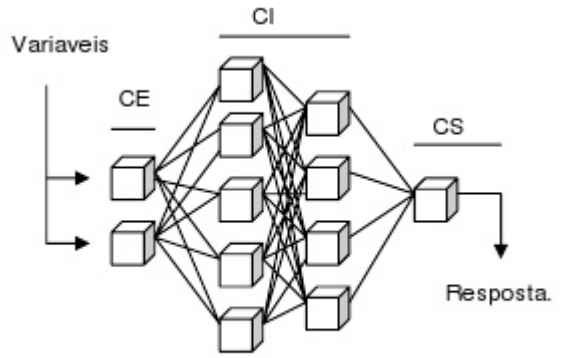


Figura 2: Rede Perceptron de Camadas Múltiplas.

Figura 11: Modelo MLP

O aprendizado da Rede Neural MLP é supervisionado. No aprendizado supervisionado, para cada padrão apresentado a rede, no momento de seu treinamento, é apresentado também o padrão de saída desejado para aquele padrão de entrada. Os pesos das conexões são, então, atualizados segundo a diferença entre o padrão produzido pela rede e o padrão desejado.

Nas Redes MLP, a propagação das ativações se dá da entrada em direção à saída (*feed-forward*), não havendo ligações entre neurônios de uma mesma camada. Por outro lado, a correção dos pesos, durante o treinamento, se dá de forma retrógrada, da camada de saída até a camada de entrada, através de um algoritmo conhecido como *Backpropagation*.

O algoritmo de treinamento supervisionado conhecido como *Backpropagation* surgiu em 1986, por ocasião da publicação do livro *Parallel Distributed Processing* (RUMELHART; MCCLELLAND, 1986). De forma resumida, o algoritmo utiliza os pares (entrada/saída desejada) para calcular o erro da camada de saída da rede e então propagá-lo para as camadas intermediárias, ajustando, na sequência, os pesos das conexões da rede.

Os ajustes nos pesos das conexões podem ocorrer de forma online (*pattern-by-pattern*), ou seja, a cada vez que o algoritmo é executado, ou por ciclo (*batch*), onde a correção ocorre após todos os padrões terem sido apresentados.

Após a entrada de um padrão na Rede, o mesmo é processado pelas camadas até atingir a saída da Rede. Durante a fase de treinamento supervisionado, o erro é dado pela equação 2.11:

$$Ep = \frac{1}{2}\sum(t_{pj} - o_{pj})^2 \tag{2.11}$$

Onde t é a saída desejada e o o valor observado na saída da Rede.

Os pesos são então reajustados, iniciando-se pela camada de saída e seguindo para as camadas intermediárias. A equação 2.12 define a forma pela qual a atualização ocorre.

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_{pj}o_{pj} \tag{2.12}$$

Podemos observar que o peso no instante t+1 é uma função do peso anterior adicionado ao valor observado na saída do neurônio no instante t, multiplicado por uma constante η (taxa de aprendizagem – learning rate) e por δ (gradiente de erro local), que pode ser calculado de duas formas distintas, dependendo da camada onde o neurônio está localizado:

- A equação 2.13 mostra o cálculo para neurônios localizados na camada de saída.

$$\delta_{pj} = ko_{pj}(1 - o_{pj})(t_{pj} - o_{pj}) \tag{2.13}$$

- A equação 2.14 mostra o cálculo para neurônios localizados nas camadas intermediárias.

$$\delta_{pj} = ko_{pj}(1 - o_{pj})\sum_{k}\delta_{pk}w_{jk} \tag{2.14}$$

Onde o somatório é dado para as k unidades na camada posterior à camada onde se situa a unidade j.

Os critérios de parada do algoritmo também podem variar, podendo até mesmo serem combinados. Os critérios mais usuais são:

- Quando um erro mínimo global estipulado para a fase de treinamento é atingido;
- Quando a variação do erro torna-se muito pequena a ponto de atingir-se uma evolução muito pequena no treinamento;
- Quando um determinado número de ciclos é completado.

3 Revisão Bibliográfica

Discutem-se, neste capitulo, outros trabalhos e pesquisas envolvendo classificação de emails.

Ion Androutsopoulos et al. (ANDROUTSOPOULOS et al., 2000b) utilizam, como Base de Dados, a PU1 [1]. A base PU1 é composta por 1099 emails, dos quais 481 são spams e 618 são hams. Os emails foram coletados da caixa de correio (*mailbox*) de um dos autores do artigo, durante vinte e dois meses. Os emails em outras línguas que não o inglês e réplicas de spams coletados em um mesmo dia foram descartados.

As palavras nos emails da base de dados foram consideradas em sua forma radical (ex. earned torna-se earn). Todos os anexos e tags HTML dos emails foram removidos. Os emails foram divididos em dez partes, uma para teste e nove para treinamento. Cada experimento foi executado por dez vezes, com os resultados publicados representando a média obtida.

Para determinação das palavras mais relevantes em cada classe, foi utilizado o método de seleção de características *Mutual Information* (MI). Como classificador, foi empregado o modelo Naive Bayesian, com um conjunto de entradas variando entre 50 e 700 características, com passos incrementais de 50 elementos.

Os autores obtiveram resultados com 95% de precisão. Contudo, quando utilizaram ponderações com custo maior para classificações incorretas de emails legítimos (falsos positivos), a precisão dos resultados caiu para patamares em torno de 90% ou menos. Assim, apesar dos resultados positivos, os autores concluem que quando um custo muito alto é atribuído a classificações incorretas de emails legítimos (falsos positivos), é mais vantajoso não utilizar o sistema antispam.

Ion Androutsopoulos et al. (ANDROUTSOPOULOS et al., 2000a) empregaram, basicamente, as mesmas técnicas utilizadas em (ANDROUTSOPOULOS et al., 2000b) para a obtenção dos resultados sobre a Base de Dados LingSpam. A Base Ling Spam é composta

[1] PU1 está disponível na seção de publicações em *http://www.iit.demokritos.gr/ ionandr*.

por 16% de emails spam. Os autores concluem o classificador Naive Bayesian é ineficiente quando o custo de falsos positivos torna-se muito alto, sobretudo quando emails spam são removidos automaticamente das caixas de correio (*mailboxes*) dos usuários. O melhor resultado é obtido quando os autores utilizam 300 características dos emails, obtendo 100% de classificações corretas para o índice "spam precision". Por outro lado, o rigor exigido para a obtenção deste melhor resultado reduz, para 63%, o percentual de classificação corretas para o índice "spam recall", ou seja, uma maior quantidade de emails spam foi classificada incorretamente como ham. Os autores afirmam por fim, que máquinas de aprendizado (*learning machines*) são necessárias para que os filtros antispam sejam capazes de se adaptarem a novos formatos de mensagens ilegítimas.

Zorkadis, Karras e Panayotou (ZORKADIS; KARRAS; PANAYOTOU, 2005) fazem uso da Base de Dados LingSpam em um estudo comparativo entre modelos de filtragem de spam, visando a redução de Falsos Positivos e Negativos. Como apresentado em (ANDROUTSOPOULOS et al., 2000b) e (ANDROUTSOPOULOS et al., 2000a) a base de dados é processada de forma a posicionar as palavras em suas formas radicais. Além disto, com base em uma lista, são eliminadas palavras pouco significativas, tais como "a", "as", "the", "for", etc. As palavras mais significativas para a classificação dos emails são selecionadas através do método estatístico MI. Para os testes a Base de Dados LingSpam foi divida em 60% dos emails para treinamento e 40% para testes, mantendo as proporções originais entre emails Spam e Ham. Foram realizados testes com 100 e 500 características, utilizando os modelos Naïve-Bayes (ELKAN, 1997), AdaBoostM1 (TRESP, 2001), Classification via Regression (TRESP, 2001), MultiBoostAB (TRESP, 2001), Random Committee (TRESP, 2001), ADTree (Alternate Decision Tree) (FREUND; MASON, 1999), ID3Tree (FRANK et al., 1998), RandomTree (TRESP, 2001) e o modelo proposto pelos autores, que combina os modelos Random Committee e ADtree (Classification via Regression), complementares quanto à eficiência na classificação de Falsos Positivos e Falsos Negativos.

Os autores concluem que o modelo com a menor incidência de Falsos Positivos foi o Random Committee, enquanto que o ADtree e o ID3Tree foram os de melhor desempenho quanto a Falsos Negativos. Os autores não apresentam seus resultados em termos percentuais de classificações corretas. Contudo, mencionam que, a partir de 500 características, os modelos propostos classificam incorretamente entre 1 e 5 emails Ham e entre 11 e 52 emails Spam.

Chuan et al. (CHUAN et al., 2005) comparam o desempenho de três modelos — rede neural perceptron multicamadas (PMC) com algoritmo de treinamento backpropagation,

least vector quantization (LVQ) e Naive Bayesian — para classificação de emails. Utilizam a Base de Dados SpamAssassin, onde são selecionados, aleatoriamente, 580 emails spam e 420 hams, e o método estatístico MI, para seleção das características mais relevantes.

Utilizaram 100 características dos emails para compor os vetores de entrada. Propõem um método composto por dois estágios. O primeiro tem a função de definir a subclasse a qual o email pertence. Os emails são divididos, então, em subclasses, de acordo com a categoria em que se inserem (ex.: promoções, compras, conteúdo adulto, etc). Nesta etapa, os três modelos são treinados para classificar os emails nestas subclasses. No segundo estágio, os modelos são treinados para classificar os emails destas subclasses em apenas duas classes — Spam e Ham. Os resultados obtidos pelos autores indicam que o modelo LVQ apresenta desempenho ligeiramente superior ao PMC, com percentuais de classificações corretas de 98% e 93% para os índices "spam precison" e "spam recall", respectivamente. O modelo PMC obtém resultados de 90% e 91% para estes índices. Os resultados obtidos pelo modelo Naive Bayesian são inferiores aos obtidos pelos outros dois modelos.

Thomas Lynam e Gordon Cormack (LYNAM; CORMACK, 2005) propõe encontrar uma situação mais próxima possível da encontrada por um sistema antispam na prática. Para tanto, constroem uma Base de Dados partindo de uma composição de emails de fontes variadas, que vão de mensagens privadas capturadas a partir do *feedbak* de usuários para um sistema antispam até emails presentes na Base de Dados SpamAssassin. A coletânea de emails passa por cinco revisões, refinando a cada passo a classificações dos emails entre Hams e Spams. Os autores definem seu trabalho como ainda em desenvolvimento e pondera que, apesar de algumas mensagens utilizadas na Base terem perdido algumas de suas características originais, sua proposta de criar a maior e mais representativa Base de Dados para testes e simulações é factível. A Base de Dados TREC, decorrente das pesquisas de Thomas Lynam e Gordon Cormack foi utilizada para validação do modelo proposto por este trabalho.

Hao Xu e Bo Yu (XU; YU, 2010) almejam obter uma evolução nos resultados de classificação de emails. Em sua metodologia, propõe o uso de um analisador léxico, que correlaciona a ocorrência de uma palavra com as demais, e uma alteração no algoritmo de treinamento *backpropagation* de uma rede neural perceptron multicamadas (PMC), a fim de otimizar o tempo de convergência e evitar mínimos locais. O trabalho utilizou a Base de Dados LingSpam, com a substituição das palavras por suas formas radicais. Mil emails, selecionados de forma aleatória, foram utilizados para treinar e testar o PMC. Os emails

foram divididos em dez partes, sendo nove utilizadas no treinamento e validação e uma parte utilizada nos testes. Os melhores resultados foram obtidos pelo PMC, empregando *backpropagtion* em sua forma revisada, em conjunto com o analisador léxico ("accuracy" acima de 98%). Os autores concluem que o *backpropagation* revisado produz o maior impacto na qualidade dos resultados, mas que o uso de um analisador léxico ponderado para as palavras também contribui para esta qualidade.

Vladimir Vapnik, Donghui Wu e Harris Drucker (VAPNIK; WU; DRUCKER, 1999) comparam uma support vector machine (SVM) com outros três modelos, para classificação de emails. São utilizadas duas bases de dados. A primeira, com 850 emails spam, 2150 emails ham, para treinamento. O campo de assunto e o próprio corpo do email são retirados e reinseridos, de modo a gerar diferentes combinações. As palavras foram ordenadas, por relevância, através do método MI. A segunda base de dados é constituída por 314 emails spam e 303 ham. Nenhum método de seleção de características foi empregado sobre esta base. Todas as palavras com menos de três ocorrências foram descartadas. As demais compuseram o vetor de características. Os autores concluem que tanto SVMs quanto Boosting Trees obtiveram bom desempenho. Quanto ao número de palavras para compor o vetor de características, conclui que o melhor desempenho dos modelos é obtido com o uso de todas as palavras da base convertidas na forma minúscula. Em seu melhor resultado, a SVM obteve taxas entre 2% e 3% de falsos alarmes positivos.

Ashutosh Deshpande e Joon Park (DESHPANDE; PARK, 2006) propõe, para a detecção de spams, o uso de um filtro híbrido composto por três técnicas - listas de detecção, verificação de conteúdo e detecção de email forjado. Listas brancas e negras são as listas de detecção mais empregadas. Nas listas brancas, são cadastrados remetentes de emails legítimos, enquanto que nas listas negras, são cadastrados remetentes de emails spam. Ambas as listas podem ser implementadas tanto no cliente de email, quanto no servidor de email.

As técnicas baseadas em verificação de conteúdo geralmente analisam o corpo e anexos do email em busca de características-chave que indiquem a procedência legitima ou não da mensagem. Tais técnicas apóiam-se, basicamente, na categorização de textos e estruturas, fazendo uso de modelos estatísticos e neurais para a classificação dos emails.

A detecção de emails forjados é realizada por meio da identificação de informações falsas inseridas nos campos de cabeçalho do email. As técnicas para inserção de informações falsas chegam a ser bem sofisticadas, atualmente, exigindo do filtro, por vezes, o rastreamento do endereço IP de origem do email.

Os autores empregam as técnicas na ordem que segue. Primeiro, é verificada a integridade do DNS do email. Em seguida, é verificado se o endereço de origem do email está presente em uma lista negra. Terceiro, são comparados o campo "from" e a informação obtida via DNS, para reduzir a ocorrência de falsos positivos. Quarto, o endereço de origem é pesquisado em uma lista branca e, por último, é aplicado um filtro de análise de conteúdo.

A base de dados utilizada é composta por 905 emails, dos quais 234 são legítimos e 671 são spams. A base foi formada por emails coletados da caixa de correio de um único usuário. O autores comparam os resultados obtidos pelo método híbrido com o uso individual de cada uma das técnicas. O método hibrido apresentou desempenho superior em relação as técnicas aplicadas individualmente, alcançando percentuais de 98% para os índices de "spam e ham precision" e percentuais de 99% e 96%, respectivamente, para os índices de "spam e ham recall". Por fim, os autores abordam a limitação de terem usado, como base de dados, emails de um único usuário. Pretendem, no futuro, estender seus estudos a outras bases, a fim de observarem o desempenho do método proposto em um ambiente mais próximo da realidade encontrada pelos filtros antispam.

Byeong Man Kim, Sin-Jae Kang e Jong-Wan Kim (KIM; KANG; KIM, 2005) propõe, para detecção de spams, o uso de lógica fuzzy para otimizar o processo de seleção de características de uma base de dados de emails. Como modelo classificador, os autores utilizam uma support vector machine (SVM). A base de dados utilizada é composta por 4792 emails, sendo 2218 hams, 1100 spams de conteúdo pornográfico, 1077 spams de conteúdo financeiro e 397 spams de conteúdo comercial. Os autores comparam os métodos estatísticos para seleção de características — Information Gain e Chi-Quadrado — com sua abordagem, que utiliza lógica fuzzy. Concluem que a utilização de lógica fuzzy na seleção de características produz resultados percentuais de 79% e 91%, respectivamente, para os índices de "spam precision" e "spam recall". Estes resultados correpondem, em média, a uma redução entre 6% a 10% do erro médio, em relação aos métodos Information Gain e Chi-Quadrado.

Sheng-Yi Chen e Chih-Chien Wang (CHEN; WANG, 2007) propõe a identificação de emails legítimos e spams através da análise do cabeçalho do email. Sugerem que a identificação de cabeçalhos suspeitos pode ser realizada através da verificação da ocorrência das principais técnicas utilizadas por spammers para ocultar a procedência do email. Por exemplo, citam que apenas 7,2% dos spams possuem endereço de destino nos campos "To" e "CC". Em sua maioria, portanto, possuem tal endereço no campo de cópia oculta "BCC".

Ademais, o campo “X-Mailer”, que indica o software cliente utilizado para enviar o email (MUA), apresenta, geralmente, informações distorcidas, uma vez que muitos spams são originados a partir de programas que automatizam o envio de mensagens. Igualmente, o campo identificador da mensagem (Message-ID), composto por duas partes separadas pelo símbolo “@”, é alterado com o intuito de ocultar o nome real do domínio originário do spam, ou do primeiro ponto de transferência pelo qual ele passou. Assim, o domínio apresentado no campo do remetente não combina com o domínio apresentado no Message-ID.

Os estudos foram realizados sobre uma base com 10.024 emails spam, coletados em um período de dois meses e disponibilizados por Spam Archive [2], além de 599 emails legítimos usuais e 635 emails legítimos comerciais solicitados, coletados por três voluntários no período de uma semana. Com base na análise das características dos emails coletados, os autores determinam três regras para considerar um email como sendo legitimo, e quatro regras para considerá-lo como spam. Os autores concluem que se o usuário utilizar uma estratégia “conservadora” de somente bloquear emails que atendam plenamente as regras, a classificação não gera falsos positivos. No entanto, o número de falsos negativos (spams que passaram pelo filtro) produzido gira em torno de 20,89%. Caso uma estratégia “agressiva” seja utilizada, o filtro é capaz de identificar 92,5% dos spams. A taxa de falsos positivos sobe, porém, de 0 para 10,28%.

[2]HTTP://spamarchive.org

4 Testes, resultados e análises

Este capítulo tem por finalidade apresentar a metodologia e os procedimentos que levaram aos resultados obtidos.

Como descrito no capítulo 2, o primeiro estágio do sistema antispam consiste na limpeza dos emails da base de dados. Neste estágio de pré-processamento, tags HTML são classificadas sob os três critérios apresentados (seção 2.1.3.1), e por conseqüência, substituídas pela saída padronizada. Ocorrem também, neste estágio, a busca por padrões conhecidos utilizados por *spammers*.

O segundo estágio consiste na utilização dos métodos de seleção de características apresentados na seção 2.2. A função destes é elencar as palavras mais significativas para determinação da classe a que o email pertence (spam ou ham). A lista de palavras é então utilizada para formar os vetores de entrada para o classificador neural. Para os testes, foram elaborados vetores com até 100 entradas (as 100 palavras mais significativas para cada método).

O terceiro estágio é o estágio de classificação. Este, consiste em entrar com os vetores elaborados no classificador neural.

Os métodos de seleção de características podem, em virtude do universo reduzido de, no máximo, 100 palavras, deixar de fora palavras presentes em um determinado email, gerando, portanto, um padrão zerado (sem ocorrência de pelo menos uma das palavras significativas).

Desta forma, é necessário gerar um conjunto reduzido de vetores, sem a presença de padrões zerados. Este conjunto é utilizado para treinamento e validação do classificador neural. Para o teste do classificador, os padrões zerados são reintroduzidos.

Além disto, os conjuntos de Treinamento, Validação e Teste são balanceados. A classe, Spam ou Ham, com menor número de padrões tem seus padrões duplicados aleatoriamente, de forma a que o número de padrões Spam e Ham seja igual em cada conjunto.

O vetor de características, representando os emails da Base de Dados sendo processada, são gerados de forma normalizada, ou seja, as características apresentadas nos vetores de entrada para o classificador neural são ponderadas pela característica com maior número de ocorrências no email, gerando valores entre 0 e 1.

Após estes procedimentos, inicia-se o treinamento e validação do classificador neural. A dimensão dos vetores de entrada varia de cinco em cinco elementos até atingir 100 elementos. Os conjuntos de Treinamento, Validação e Teste são divididos segundo a proporção de 40%, 20% e 40% respectivamente, sendo que o conjunto de testes incorpora também, como mencionado, os padrões "zerados" descartados para o treinamento e validação.

4.1 Bases de Dados Utilizadas

Os estudos realizados para a validação do sistema antispam utilizaram três bases de dados distintas, todas públicas, com emails reais e disponíveis na Internet.

- **Ling-Spam Corpus** – a Base de Dados é composta por 2893 emails, sendo destes 481 spams e 2412 hams. (ANDROUTSOPOULOS et al., 2000a) descreve em seu trabalho a Base de Dados Ling-Spam como sendo uma composição de mensagens spam, recebidas a partir de uma caixa de emails pessoal, com mensagens legítimas, enviadas a uma lista de discussão. As principais deficiências da base de dados em questão residem no fato da amostra de Spam ser baseada em um único perfil de usuário, além do fato de que os anexos e tags HTML terem sido removidos, tornando a tarefa de classificação mais simples.

- **SpamAssassin Corpus** – a base de dados SpamAssassin (SPAMASSASSIN, 2008) é dividida em cinco partes — dois diretórios com emails spam e três com emails ham — em um total de 6037 mensagens, sendo 1844 spams e 4193 hams. Os cabeçalhos originais dos emails foram preservados, com algumas alterações somente no endereço por questões de privacidade. Por apresentar emails recolhidos de fontes variadas, onde o conteúdo original foi preservado (incluindo conteúdos HTML), a base de dados SpamAssassin constitui um conjunto de testes desafiador, que tenta satisfazer a heterogeneidade encontrada por sistemas antispam em uma situação real.

- **Trec Corpus** – A base de dados Trec (TREC, 2007) possui 75.288 emails, sendo 25.214 hams e 50.074 spams, coletados entre 8/04/2007 e 06/07/2007 de um servidor

particular. A principal característica da base de dados Trec é a presença de spams em imagens, identificadas no estágio de pré-processamento pela Tag img, que, como definido na seção 2.1.3.1, é processada integralmente.

4.2 Medidas de Erro e Desempenho

Para fins estatísticos, cada estudo foi realizado por dez vezes e, em cada uma delas, o classificador neural foi treinado com pesos iniciais diferentes, gerados aleatoriamente. Portanto, para cada estudo, o desempenho do classificador neural é aferido através da média simples dos erros dos dez resultados na classificação de emails ham e spam, bem como da dispersão destes erros em relação à média. Para cada estudo são reportados:

- As características dos conjuntos de Treinamento, Validação e Teste, apresentando-se o número de padrões utilizados em cada um, bem como o número de padrões zerados e duplicados;
- Os resultados obtidos, separados pelos métodos de seleção de características (DF, MI e Chi-Quadrado), quando utilizados;
- O tempo gasto em processamento para treinamento e teste do classificador neural.

Em todos os estudos, utilizou-se um computador com as configurações de hardware e de software exibidas na tabela 3:

Tabela 3: Configuração da Máquina utilizada nos testes

Modelo	VAIO VGN-FS960
Processador	Intel(R) Pentium(R) M processor 1.73GHz
Memória	1,50 GB
Sistema Operacional	Windows 7 Ultimate 32-bit
Ambiente Pré-Filtro	Java 6 Update 17 - Sun Microsystems
Ambiente Rede Neural	Matlab R2008a

4.3 Teste utilizando a Base de Dados SpamAssassin

Como descrito na seção anterior, a Base SpamAssassin tenta reproduzir um ambiente mais próximo possível da realidade encontrada por um usuário convencional, por este motivo ela foi escolhida como base de emails na maior parte dos estudos.

Para comprovar o efeito do pré-processamento e da seleção de características no desempenho do sistema antispam, foram realizados quatro estudos com a Base SpamAssassin.

4.3.1 Estudo 1

Para o Estudo 1, não foram utilizados o pré-processamento nem os métodos de seleção de características. A tabela 4 apresenta as características relacionadas aos conjuntos de Treinamento, Validação e Teste utilizados no estudo. A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice B.

Tabela 4: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 1

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	81	40	0	0	0	41	162
Validação	41	20	0	0	0	21	82
Teste	81	39	4135	1600	0	42	5897

Como pode-se verificar, por não se utilizar nenhum método para seleção das palavras mais significativas, uma grande quantidade de padrões zerados é gerada e descartada, reduzindo os conjuntos de Treinamento e Validação a apenas 162 e 82 padrões, respectivamente. No conjunto de Teste, como mencionado, os padrões zerados são reinseridos.

A tabela 5 traz os resultados obtidos no Estudo 1 e a tabela 6 mostra os tempos gastos para Treinamento e Teste do classificador neural

Tabela 5: Resultados obtidos pelo Estudo 1

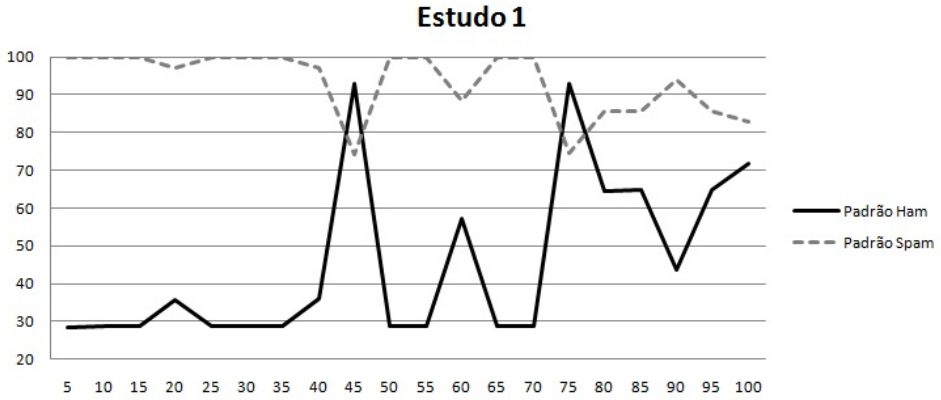
Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	28,5583	±	0,0000	100,0000	±	0,0000
2	10	28,5767	±	0,0000	100,0000	±	0,0000
3	15	28,5767	±	0,0000	100,0000	±	0,0000
4	20	35,7614	±	7,1294	97,0262	±	2,8263
5	25	28,632	±	0,0000	99,8525	±	0,0000
6	30	28,8164	±	0,0000	99,8156	±	0,0000
7	35	28,8164	±	0,0000	99,8156	±	0,0000
8	40	35,9347	±	7,1184	96,9875	±	2,8282
9	45	92,878	±	7,1221	74,3621	±	2,8282
10	50	28,8496	±	0,0037	99,8304	±	0,0147
11	55	28,8532	±	0,0000	99,8156	±	0,0000
12	60	57,3119	±	28,4587	88,5029	±	11,3127
13	65	28,8901	±	0,0000	99,7419	±	0,0000
14	70	28,9049	±	0,0147	99,8304	±	0,0590
15	75	92,8319	±	7,1313	74,6958	±	2,8853
16	80	64,4856	±	35,4775	85,8075	±	13,9971
17	85	64,6921	±	35,1752	85,6213	±	13,7703
18	90	43,514	±	14,6608	93,9012	±	5,9882
19	95	64,7824	±	35,1364	85,6895	±	13,7021
20	100	71,8215	±	28,1416	82,9130	±	10,9919

Tabela 6: Tempos em segundos gastos no Treinamento e Teste do Estudo 1

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	0,8251	0,0717
10	0,9236	0,0656
15	0,7674	0,0749
20	0,7267	0,0873
25	0,9891	0,0828
30	0,9062	0,0843
35	1,2689	0,0922
40	0,9422	0,1344
45	1,0314	0,0999
50	1,0293	0,0913
55	0,8985	0,0967
60	0,9110	0,0955
65	1,0265	0,0969
70	0,7578	0,1000
75	0,7454	0,1016
80	0,7878	0,1044
85	0,7595	0,0982
90	0,7143	0,0982
95	0,8717	0,1391
100	0,7251	0,1064

Ainda que os resultados para classificação de Spams indiquem valores extremamente positivos, principalmente com vetores de entrada com uma quantidade menor de elementos, verifica-se uma alta incidência de falsos positivos (hams classificados como spams).

Figura 12: Desempenho comparativo de classificações corretas dos padrões spam e ham



A baixa qualidade da classificação apresentada deve-se, principalmente, ao fato das palavras chave escolhidas para compor os elementos do vetor de entrada não serem representativas. O erro comporta-se de forma absolutamente aleatória, não apresentando redução com o aumento da dimensionalidade do vetor de entrada, como mostra o gráfico da figura 12. Outro ponto negativo pode ser observado em relação à dispersão das medidas, apresentando variações acentuadas.

Por último, é importante ressaltar que a comparação do tempo de Treinamento do classificador neural em relação aos demais estudos que vão se seguir não é pertinente, devido à baixa quantidade de padrões existente nos conjuntos de Treinamento e Validação.

4.3.2 Estudo 2

No Estudo 2, foi utilizado o pré-processamento dos emails, porém nenhum método de seleção de características foi empregado. A tabela 7 apresenta as características relacionadas aos conjuntos de Treinamento, Validação e Teste utilizadas no estudo. A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice C.

As características mostram que, apesar do pré-processamento reduzir, em parte, o universo de palavras chave que formarão os vetores de entrada, a não utilização de um método de seleção de características faz com que a escolha das palavras chave não seja eficiente, gerando um número significativo de padrões zerados. Assim, os conjuntos utilizadas para Treinamento e Validação ficaram reduzidas a 780 e 390 padrões, respectivamente.

Tabela 7: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 2

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	390	170	0	0	0	220	780
Validação	195	85	0	0	0	110	390
Teste	389	169	3219	1420	0	220	5417

A tabela 8 apresenta os resultados obtidos no Estudo 2, e a tabela 9 mostra os tempos gastos para Treinamento e Teste do classificador neural.

Tabela 8: Resultados obtidos pelo Estudo 2

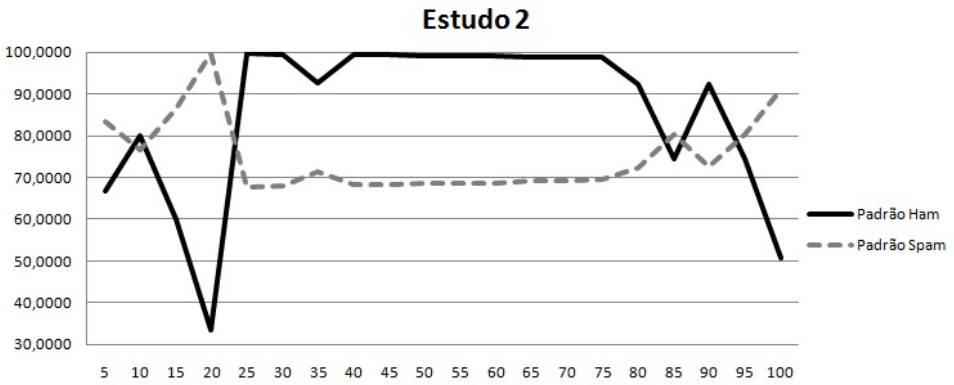
Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	66,6882	±	33,2933	83,3026	±	16,6974
2	10	80,0166	±	19,9649	76,6236	±	10,0185
3	15	60,0517	±	26,6199	86,5867	±	13,3210
4	20	33,4318	±	0,0000	99,9077	±	0,0000
5	25	99,6345	±	0,0037	67,6888	±	0,0240
6	30	99,4185	±	0,0092	68,0801	±	0,0351
7	35	92,8078	±	6,6199	71,2701	±	3,3173
8	40	99,3797	±	0,0258	68,1650	±	0,1569
9	45	99,3668	±	0,1237	68,2278	±	0,3157
10	50	99,2579	±	0,0517	68,5066	±	0,0738
11	55	99,2579	±	0,0701	68,4918	±	0,1809
12	60	99,2394	±	0,0517	68,6782	±	0,0314
13	65	98,7576	±	0,0609	69,1637	±	0,1403
14	70	98,7318	±	0,0498	69,0991	±	0,0757
15	75	98,7299	±	0,0960	69,6123	±	0,2049
16	80	92,4165	±	6,0698	72,4497	±	2,6878
17	85	74,5265	±	24,1813	80,3932	±	10,8713
18	90	92,4183	±	6,2156	72,5272	±	2,8023
19	95	74,5099	±	23,9579	80,4412	±	10,6424
20	100	50,7735	±	11,9697	91,0338	±	5,3849

Tabela 9: Tempos em segundos gastos no Treinamento e Teste do Estudo 2

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	1,1406	0,0642
10	1,1484	0,0640
15	1,2077	0,0719
20	1,4311	0,0734
25	1,0829	0,0734
30	1,1343	0,0734
35	1,5296	0,0764
40	1,3547	0,0782
45	1,2049	0,0842
50	1,0608	0,0845
55	1,2532	0,0876
60	1,3922	0,0907
65	2,0327	0,0924
70	1,7360	0,0937
75	2,0125	0,0938
80	1,2454	0,0985
85	1,3797	0,1000
90	1,4267	0,1032
95	1,4359	0,1108
100	1,4345	0,1094

Os resultados obtidos no Estudo 2 são melhores do que os do Estudo 1, embora, como apresentado no gráfico 13, o comportamento do erro permaneça aleatório, não apresentando redução em relação ao crescimento da dimensionalidade do vetor de entrada. Além disto, a dispersão das medidas também permanece significativa.

Figura 13: Desempenho comparativo de classificações corretas dos padrões spam e ham



Novamente, tal como no Estudo 1, a comparação do tempo de Treinamento do classificador neural em relação aos demais estudos que vão se seguir não é pertinente, devido à baixa quantidade de padrões existente nos conjuntos de Treinamento e Validação.

4.3.3 Estudo 3

No Estudo 3, o pré-processamento dos emails não foi realizado, mas foram empregados os três métodos de seleção de características. Os resultados são apresentados em três subseções, uma para cada um dos métodos utilizados — DF, Chi-Quadrado e MI. A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice D.

4.3.3.1 Estudo 3: Método DF

A tabela 10 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizados no Estudo 3, para o método DF.

Tabela 10: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 3 - Método DF

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1659	724	0	0	0	935	3318
Validação	829	362	0	0	0	467	1658
Teste	1659	723	46	35	0	936	3399

Apenas observando-se as características dos conjuntos, já se pode perceber a diferença em relação aos estudos anteriores. Desta vez, o número de padrões zerados reduziu-se significativamente, reforçando a hipótese de que a utilização de métodos de seleção de características é fundamental para uma boa construção dos conjuntos de dados.

A tabela 11 apresenta os resultados obtidos no Estudo 3, para o método DF, e a tabela 12, os tempos gastos para Treinamento e Teste do classificador neural.

Tabela 11: Resultados obtidos pelo Estudo 3 - DF

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	79,02324	±	1,32392	85,63989	±	1,738746
2	10	84,97794	±	1,218006	89,43219	±	1,182705
3	15	84,64548	±	1,209183	89,80877	±	1,218003
4	20	88,84084	±	1,697558	90,50897	±	1,253308
5	25	88,54957	±	1,376873	91,58282	±	1,003233
6	30	90,45308	±	1,721095	90,65902	±	1,573988
7	35	93,11857	±	1,203295	91,3504	±	0,853187
8	40	93,13327	±	2,224183	91,13857	±	1,612239
9	45	92,78317	±	1,097379	92,39188	±	0,570761
10	50	93,0303	±	1,085617	91,64166	±	1,11503
11	55	92,41247	±	1,703447	92,06237	±	1,935858
12	60	92,02118	±	2,712567	93,18917	±	1,809353
13	65	91,2592	±	3,680495	93,43336	±	1,241538
14	70	92,83613	±	1,956462	93,4569	±	1,282729
15	75	92,70962	±	2,230068	93,45101	±	0,982644
16	80	93,33627	±	2,191824	93,47749	±	1,509264
17	85	93,18329	±	2,491911	93,46572	±	1,732864
18	90	93,85702	±	1,906444	93,49515	±	0,820826
19	95	94,10415	±	1,5181	93,55987	±	1,115029
20	100	93,39806	±	2,306562	93,89821	±	0,812006

Tabela 12: Tempos em segundos gastos no Treinamento e Teste do Estudo 3 - DF

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	6,0406	0,0501
10	4,1439	0,0484
15	3,5514	0,0533
20	3,4641	0,0611
25	4,4016	0,0577
30	3,7423	0,0592
35	4,0610	0,0576
40	3,9280	0,0642
45	3,7172	0,0624
50	3,7939	0,0670
55	3,2281	0,0642
60	3,4702	0,0643
65	3,6502	0,0701
70	3,8282	0,0687
75	3,8908	0,0749
80	3,6091	0,0748
85	3,6891	0,0750
90	3,7391	0,0796
95	3,9452	0,0796
100	3,9453	0,0782

Com uma melhor construção dos conjuntos de dados, os resultados também apresentaram um comportamento bem melhor, dentro do esperado, com o erro de classificação decaindo à medida em que o número de elementos do vetor de entrada aumentava.

O desempenho na classificação de Spam mostrou-se ligeiramente superior ao da classificação de hams, inclusive com uma dispersão menor que a apresentada na classe ham.

Um fato relevante pode ser observado nos tempos gastos no treinamento do classificador neural. Com a maior dimensionalidade do vetor de entrada, os tempos gastos em treinamento deveriam, teoricamente, aumentar, devido ao aumento do número de neurônios da camada de entrada e do consequente aumento no número de conexões na arquitetura do classificador neural. No entanto, os tempos inicialmente decaem e, posteriormente, estabilizam-se, sugerindo que, apesar da maior carga de processamento computacional por época de treinamento, a convergência do treinamento melhora, reduzindo-se o número de épocas de treinamento necessárias.

Em relação ao tempos gastos no teste do classificador neural, ocorre o comportamento esperado. Quanto maior a dimensionalidade do vetor de entrada, maior o tempo que o classificador neural leva para produzir a resposta.

4.3.3.2 Estudo 3: Método Chi-Quadrado

A tabela 13 apresenta as características relacionadas aos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 3, para o método Chi-Quadrado.

Tabela 13: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 3 - Método Chi-Quadrado

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1467	716	0	0	0	751	2934
Validação	733	358	0	0	0	375	1466
Teste	1467	716	526	54	0	751	3514

Apesar das características para o método Chi-Quadrado serem bem superiores às obtidas nos Estudos 1 e 2, é possível perceber um aumento no número de padrões zerados em relação ao número produzido pelo método DF, seção 4.3.3.1, sugerindo, assim, que a escolha das características mais significativas, feita pelo método DF, para compor os elementos dos vetores de entrada, foi mais acurada e abrangente para o conjunto de dados estudado. Esta condição, por si só, não implica, porém, em uma maior eficiência do

método DF, pois exige-se, igualmente, que os resultados produzidos pela aplicação deste método sejam também superiores aos do método Chi-Quadrado.

A tabela 14 apresenta os resultados obtidos no Estudo 3, para o método Chi-Quadrado, e a tabela 15, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 14: Resultados obtidos pelo Estudo 3 - Chi-Quadrado

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	94,1377	±	1,1668	86,0387	±	1,1497
2	10	97,0233	±	0,6716	86,9636	±	0,7086
3	15	94,9004	±	1,0586	89,9260	±	1,0529
4	20	94,8976	±	1,2038	90,3529	±	1,5367
5	25	95,6147	±	1,4485	90,2504	±	0,8993
6	30	95,3273	±	0,3472	90,1110	±	0,8680
7	35	95,0000	±	0,8737	90,4041	±	0,8196
8	40	94,9801	±	0,5293	92,5982	±	1,5680
9	45	93,3608	±	3,1104	92,6409	±	1,2920
10	50	95,5265	±	0,8196	91,8241	±	0,8623
11	55	95,3870	±	0,7427	92,5270	±	1,1497
12	60	95,5265	±	0,7171	92,3734	±	0,7684
13	65	95,6175	±	0,5692	92,5242	±	0,9875
14	70	95,6887	±	0,5265	92,0803	±	0,9476
15	75	95,5208	±	0,7513	92,3847	±	0,6944
16	80	95,7655	±	0,6773	92,5953	±	1,0472
17	85	95,7826	±	0,5464	92,5014	±	0,9249
18	90	95,0199	±	0,9960	92,8201	±	0,7314
19	95	95,4183	±	0,3700	92,8458	±	0,8139
20	100	95,6061	±	0,5009	92,3961	±	1,0871

Tabela 15: Tempos em segundos gastos no Treinamento e Teste do Estudo 3 - Chi-Quadrado

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	3,6156	0,0516
10	3,3561	0,0578
15	3,2703	0,0532
20	3,0313	0,0564
25	3,9250	0,0610
30	3,3296	0,0579
35	3,4236	0,0564
40	2,9204	0,0641
45	3,1531	0,0625
50	2,7360	0,0670
55	2,9765	0,0672
60	3,1299	0,0671
65	2,8375	0,0717
70	3,0126	0,0733
75	3,0922	0,0720
80	3,0735	0,0797
85	3,2797	0,0733
90	3,3123	0,0752
95	3,2610	0,0811
100	3,3270	0,0780

Os resultados obtidos mostram que o método Chi-Quadrado, apesar de apresentar um desempenho inferior ao apresentado pelo método DF na classificação correta de emails Spam, apresentou um desempenho melhor em relação aos falsos positivos, apresentando, portanto, uma maior taxa de classificação correta de emails Ham.

Os tempos de treinamento do classificador neural são inferiores aos obtidos no estudo anterior (seção 4.3.3.1), com o método DF, porque a quantidade de padrões zerados gerada pelo método Chi-Quadrado é maior.

4.3.3.3 Estudo 3: Método MI

A tabela 16 apresenta as características relacionadas aos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 3, para o método MI.

Pela tabela 16, pode-se observar que, no Estudo 3, o número de padrões zerados produzido pelo método MI é inferior ao número produzido pelo método Chi-Quadrado. É, porém, superior ao número de padrões zerados produzido pelo método DF.

A tabela 17 apresenta os resultados obtidos no Estudo 3, para o método MI, e a tabela 18, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 16: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 3 - Método MI

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1609	719	0	0	0	890	3218
Validação	804	359	0	0	0	445	1608
Teste	1609	719	171	47	0	890	3436

Tabela 17: Resultados obtidos pelo Estudo 3 - MI

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	90,3987	±	0,1775	82,3108	±	0,2561
2	10	91,4232	±	0,3871	88,2916	±	0,2241
3	15	91,4406	±	0,5792	88,2334	±	0,7189
4	20	94,1444	±	0,6345	88,9814	±	0,5064
5	25	93,8562	±	0,7305	88,9581	±	0,5413
6	30	93,8184	±	1,3795	90,7974	±	0,8673
7	35	93,3818	±	1,1176	91,1641	±	0,8848
8	40	94,1822	±	0,4919	91,5920	±	0,7305
9	45	91,0710	±	2,5669	93,9785	±	1,2776
10	50	91,7229	±	3,0966	94,3219	±	1,5454
11	55	92,8609	±	2,5407	93,6583	±	1,5454
12	60	92,7998	±	3,0384	94,0163	±	1,3504
13	65	93,2654	±	1,9325	94,0105	±	1,4901
14	70	95,0698	±	0,5995	93,3993	±	0,9662
15	75	93,1490	±	2,1944	94,7701	±	0,8818
16	80	95,3987	±	0,2939	94,1938	±	0,6839
17	85	94,6013	±	2,1391	94,0716	±	1,0565
18	90	95,5879	±	0,4831	93,8184	±	0,5413
19	95	95,7159	±	0,6170	93,9261	±	0,3987
20	100	95,7101	±	0,4191	93,9173	±	0,2619

Tabela 18: Tempos em segundos gastos no Treinamento e Teste do Estudo 3 - MI

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	3,1639	0,0625
10	3,5564	0,0639
15	3,4435	0,0628
20	4,5967	0,0579
25	4,3624	0,0625
30	4,4235	0,0624
35	4,3217	0,0657
40	3,2454	0,0595
45	3,1440	0,0623
50	3,1468	0,0642
55	3,1470	0,0655
60	3,1064	0,0639
65	3,3547	0,0718
70	4,1108	0,0705
75	3,2702	0,0734
80	3,2033	0,0733
85	3,7359	0,0718
90	4,0312	0,0813
95	3,5658	0,0796
100	3,6749	0,0829

Analisando os resultados, percebe-se que, independentemente das características dos conjuntos de dados, o método MI, dentre os três métodos de seleção de características, é o que apresenta o melhor desempenho, tanto na classificação correta de emails ham quanto de spam.

Os tempos gastos no treinamento do classificador neural apresentam-se estáveis, sem tendências, independentemente da dimensionalidade do vetor de entrada. Isto é devido ao fato de que as maiores dimensionalidades do vetor de entrada, que acarretam maiores gastos de tempo de processamento, são compensadas por melhores convergências no treinamento, que acarretam menores gastos de tempo, em épocas, de treinamento necessários.

Em relação ao tempos gastos no teste do classificador neural, ocorre o comportamento esperado. Quanto maior a dimensionalidade do vetor de entrada, maior o tempo que o classificador neural leva para produzir a resposta.

4.3.4 Estudo 4

Verificamos, através dos resultados dos Estudos 1, seção 4.3.1, 2, seção 4.3.2, e 3, seção 4.3.3, que a utilização dos métodos de seleção de características é fundamental para a classificação correta dos padrões Spam e Ham. Além disto, apesar dos bons resultados obtidos no Estudo 3, o Estudo 2 sugere que os mesmos podem ainda ser melhorados através do uso das técnicas de pré-processamento.

O Estudo 4 tem por objetivo, portanto, avaliar os resultados na classificação de emails Spam e Ham da Base SpamAssassin, empregando tanto as técnicas de pré-processamento quanto os métodos de seleção de características (DF, Chi-Quadrado e MI).A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice E.

4.3.4.1 Estudo 4: Método DF

A tabela 19 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 4, para o método DF.

Tabela 19: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 4 - Método DF

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1677	736	0	0	0	941	3354
Validação	839	368	0	0	0	471	1678
Teste	1677	737	0	3	0	940	3357

Pela tabela 19, verifica-se que a utilização tanto do pré-processamento quanto dos métodos de seleção de características produz apenas três padrões zerados, um baixo número quando comparado ao número produzido pelos estudos anteriores.

A tabela 20 apresenta os resultados obtidos no Estudo 4, para o método DF, e a tabela 21, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 20: Resultados obtidos pelo Estudo 4 - DF

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	92,0435	±	1,7009	88,1859	±	1,3226
2	10	93,4257	±	1,5401	89,1302	±	1,4269
3	15	94,4027	±	0,7715	89,2076	±	1,5103
4	20	95,5079	±	0,8877	91,5877	±	1,5073
5	25	94,4861	±	0,6822	92,5499	±	1,2839
6	30	94,3730	±	1,9631	93,5180	±	0,9652
7	35	95,5734	±	1,3286	94,6113	±	0,6881
8	40	95,4304	±	1,5311	94,5904	±	1,0605
9	45	95,0521	±	1,6354	95,1236	±	0,7954
10	50	95,8862	±	1,0754	95,4424	±	0,7149
11	55	96,0769	±	1,2898	95,0670	±	0,6673
12	60	95,8088	±	1,7486	95,4930	±	0,5749
13	65	96,6607	±	0,6881	94,8257	±	0,5868
14	70	96,6071	±	0,6583	94,7304	±	0,6226
15	75	97,0033	±	0,7030	94,6947	±	0,7715
16	80	97,1790	±	0,4677	95,6002	±	0,8847
17	85	97,4173	±	0,5570	95,4811	±	1,0337
18	90	97,1314	±	0,7060	95,9041	±	1,1469
19	95	97,0837	±	0,8013	96,2586	±	0,8520
20	100	96,8365	±	1,0486	96,6637	±	1,0724

Tabela 21: Tempos em segundos gastos no Treinamento e Teste do Estudo 4 - DF

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	6,1576	0,0501
10	2,8094	0,0562
15	3,5436	0,0549
20	4,0890	0,0565
25	3,9063	0,0610
30	3,9970	0,0577
35	4,2702	0,0564
40	4,0529	0,0641
45	3,4812	0,0641
50	4,1485	0,0656
55	3,9328	0,0626
60	3,8140	0,0658
65	4,0250	0,0623
70	3,9203	0,0734
75	4,2264	0,0750
80	4,1422	0,0701
85	4,9859	0,0736
90	4,0281	0,0764
95	4,2764	0,0767
100	4,3202	0,0766

Os resultados mais acurados em relação aos do Estudo 3, com o método DF, comprovam a hipótese de que a utilização tanto do pré-processamento quanto dos métodos de seleção de características reduz significativamente o número de classificações incorretas de Spam e Ham. É importante ressaltar que, além do erro médio, as dispersões das medidas também decresceram.

Os tempos gastos no treinamento do classificador neural permanecem praticamente constantes. Tal como no Estudo 3 com o método MI, seção 4.3.3.3, maiores dimensionalidades do vetor de entrada são compensadas por melhores convergências do algoritmo de treinamento. Os tempos gastos no teste do classificador neural comportam-se como esperado. Quanto maior a dimensionalidade do vetor de entrada, maior o tempo que o classificador neural leva para produzir a resposta.

4.3.4.2 Estudo 4: Método Chi-Quadrado

A tabela 22 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 4, para o método Chi-Quadrado.

Tabela 22: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 4 - Método Chi-Quadrado

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1668	734	0	0	0	934	3336
Validação	834	367	0	0	0	467	1668
Teste	1667	735	24	8	0	932	3366

Tal como no Estudo 3 com o método Chi-Quadrado, seção 4.3.3.2, percebe-se um aumento no número de padrões zerados em relação ao número produzido pelo método DF, seção 4.3.4.1. Isto sugere que a escolha das características mais significativas, feita pelo método DF, para compor os elementos dos vetores de entrada, foi mais acurada e abrangente para o conjunto de dados estudado.

A tabela 23 apresenta os resultados obtidos no Estudo 4, para o método Chi-Quadrado, e a tabela 24, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 23: Resultados obtidos pelo Estudo 4 - Chi-Quadrado

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	96,5300	±	0,3803	90,2971	±	0,6120
2	10	96,1943	±	1,5478	91,9667	±	0,9091
3	15	96,1765	±	0,9418	93,4225	±	1,2953
4	20	96,6459	±	0,9180	93,6601	±	0,5169
5	25	96,6607	±	1,2656	93,4195	±	0,4605
6	30	96,6251	±	0,5526	94,8128	±	0,4635
7	35	96,6221	±	0,8705	95,0416	±	0,6209
8	40	97,3767	±	0,5140	94,9643	±	0,7279
9	45	97,3648	±	0,7338	95,0951	±	0,3892
10	50	97,7986	±	0,3001	95,1188	±	0,2288
11	55	97,8253	±	0,2911	94,9792	±	0,5348
12	60	98,5591	±	0,3417	94,8039	±	0,2941
13	65	98,4522	±	0,2941	94,9317	±	0,2258
14	70	98,0392	±	0,5645	95,2822	±	0,7784
15	75	97,8520	±	0,6922	95,2406	±	0,6179
16	80	98,0273	±	1,1111	94,8841	±	1,0755
17	85	98,1402	±	0,7605	94,9614	±	0,6655
18	90	98,3363	±	0,3268	94,7534	±	0,3981
19	95	98,1967	±	0,6150	94,8901	±	0,6536
20	100	98,2947	±	0,6358	94,9228	±	0,4130

Tabela 24: Tempos em segundos gastos no Treinamento e Teste do Estudo 4 - Chi-Quadrado

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	3,3017	0,0484
10	3,8189	0,0514
15	4,0204	0,0563
20	4,3516	0,0548
25	3,8515	0,0610
30	3,3315	0,0528
35	2,8093	0,0625
40	3,1343	0,0609
45	3,1531	0,0609
50	3,2591	0,0611
55	2,9905	0,0641
60	3,4872	0,0644
65	3,3047	0,0672
70	3,3953	0,0657
75	3,2346	0,0671
80	3,3298	0,0734
85	3,6125	0,0766
90	3,5563	0,0719
95	3,7031	0,0781
100	3,7546	0,0765

O comportamento dos resultados é similar ao apresentado pelo estudo anterior, seção 4.3.4.1. Embora, em média, o método Chi-Quadrado tenha produzido um erro superior ao do método DF, na classificação de Spam, apresentou, em média, um erro inferior na classificação de Ham.

Os tempos de treinamento do classificador neural são inferiores aos obtidos no estudo anterior (seção 4.3.4.1), com o método DF, porque a quantidade de padrões zerados gerada pelo método Chi-Quadrado é maior.

4.3.4.3 Estudo 4: Método MI

A tabela 25 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 4, para o método MI.

Tabela 25: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 4 - Método MI

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	1677	736	0	0	0	941	3354
Validação	838	368	0	0	0	470	1676
Teste	1677	736	1	4	0	941	3359

Pela tabela 25, verifica-se que a utilização tanto do pré-processamento quanto do método MI produz apenas cinco padrões zerados. um baixo número, portanto.

A tabela 26 apresenta os resultados obtidos no Estudo 4, para o método MI, e a tabela 27, os tempos gastos no Treinamento e Teste do classificador neural.

Os resultados indicam que o método MI com utilização de pré-processamento apresenta desempenho superior a ambos os métodos DF e Chi-Quadrado também com utilização de pré-processamento.

Tabela 26: Resultados obtidos pelo Estudo 4 - MI

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	93,1561	±	0,1963	87,7366	±	0,5888
2	10	96,2235	±	1,1808	91,1871	±	0,8041
3	15	96,1349	±	1,3960	92,1779	±	1,2947
4	20	97,2903	±	0,3039	94,2513	±	1,0003
5	25	96,9073	±	0,9718	94,5837	±	1,1681
6	30	96,7205	±	0,8040	95,2707	±	0,6774
7	35	96,4704	±	0,7756	95,4796	±	1,3042
8	40	96,4039	±	0,9687	95,6252	±	1,0700
9	45	96,8534	±	0,3925	95,3340	±	0,9053
10	50	96,7553	±	0,9971	95,5081	±	0,9782
11	55	96,7996	±	0,7566	95,6695	±	0,3102
12	60	97,6638	±	0,4368	95,8405	±	0,7091
13	65	98,1861	±	0,6869	95,8278	±	0,9750
14	70	98,8161	±	0,3925	95,5049	±	0,9813
15	75	98,8382	±	0,4970	95,7297	±	0,7249
16	80	98,9459	±	0,4020	95,9038	±	0,7407
17	85	98,7686	±	0,3134	96,1475	±	0,8642
18	90	98,8984	±	0,4052	95,9797	±	0,9813
19	95	98,8889	±	0,5413	96,2963	±	0,9180
20	100	98,8699	±	0,3704	96,1475	±	0,6236

Tabela 27: Tempos em segundos gastos no Treinamento e Teste do Estudo 4 - MI

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	4,4765	0,0498
10	3,0500	0,0501
15	2,7186	0,0531
20	4,8314	0,0562
25	3,0234	0,0532
30	3,1267	0,0530
35	3,3015	0,0562
40	3,2938	0,0562
45	3,3326	0,0548
50	3,3719	0,0578
55	3,5064	0,0609
60	3,3095	0,0641
65	3,5627	0,0624
70	3,5345	0,0624
75	3,7860	0,0717
80	3,5517	0,0671
85	3,6735	0,0670
90	3,5717	0,0702
95	3,5780	0,0704
100	3,6405	0,0673

4.3.5 Análise dos Resultados sobre a Base SpamAssassin (Estudos 1 a 4)

Os resultados obtidos nos Estudos 1 e 2 são pobres quando comparados aos dos Estudos 3 e 4. Isto sugere que os métodos de seleção de características são importantes para a correta classificação dos emails realizada pelo classificador neural.

Os gráficos das figuras 14 e 15 apresentam os percentuais de classificações corretas de Spam e Ham obtidas pelos três métodos no Estudo 3, variando com o número de características selecionadas.

Figura 14: Comparativo Classificação Hams

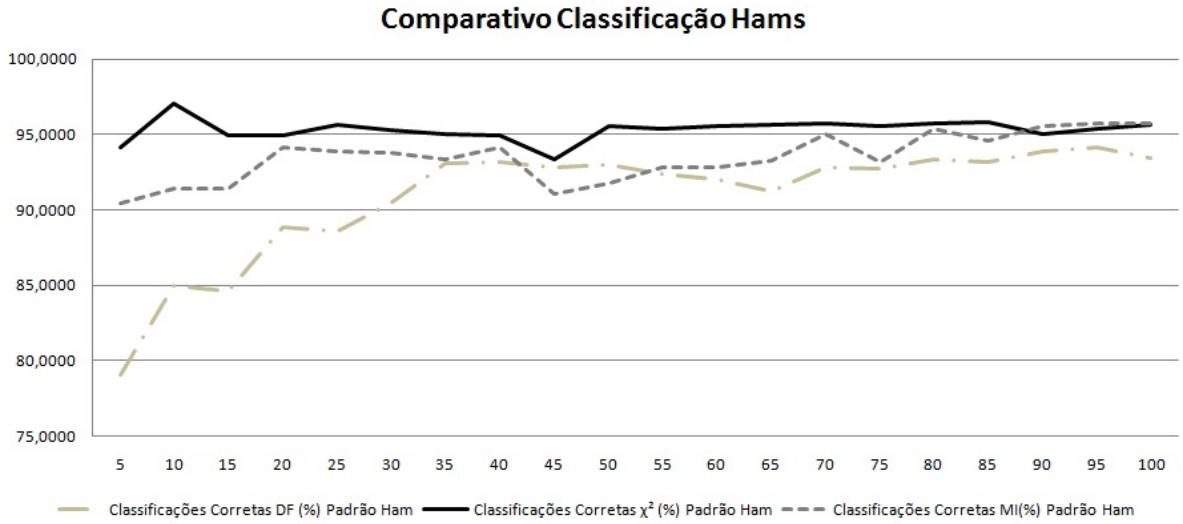
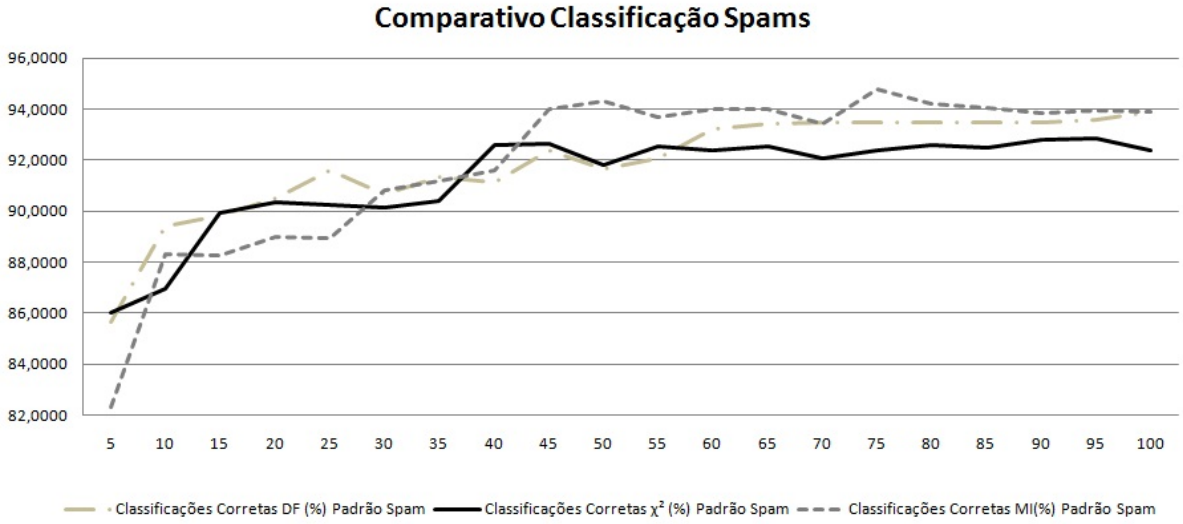


Figura 15: Comparativo Classificação Spams



Analisando-se os gráficos, verifica-se que:

- Enquanto a classificação de Hams apresenta um pequeno ganho com o crescimento do número de características selecionadas, a classificação de Spams, por sua vez, apresenta um grande ganho com tal crescimento;
- A partir de um número em torno de 50 características, o erro tende a se estabilizar, tanto para Hams quanto para Spams;
- Dentre os 3 métodos, o método Chi-Quadrado é o que apresenta o comportamento mais homogêneo, ou seja, é o que apresenta a menor evolução do erro com o aumento do número de características selecionadas;
- De modo geral, considerando-se ambas as classificações de Spams e Hams, o método MI é o que apresenta o melhor desempenho. Na classificação de Hams, porém, o método Chi-Quadrado se mostrou melhor para vetores de entrada com até 85 características;
- Para um número reduzido de entradas, o método MI é o de pior desempenho para classificação de Spam;
- Todos os métodos apresentam desempenho razoável para o problema dos falsos positivos (classificação incorreta de emails hams).

Os gráficos das figuras 16 e 17 apresentam os percentuais de classificações corretas de Spam e Ham obtidas pelos três métodos no Estudo 4, variando com o número de características selecionadas.

Figura 16: Comparativo Classificação Hams

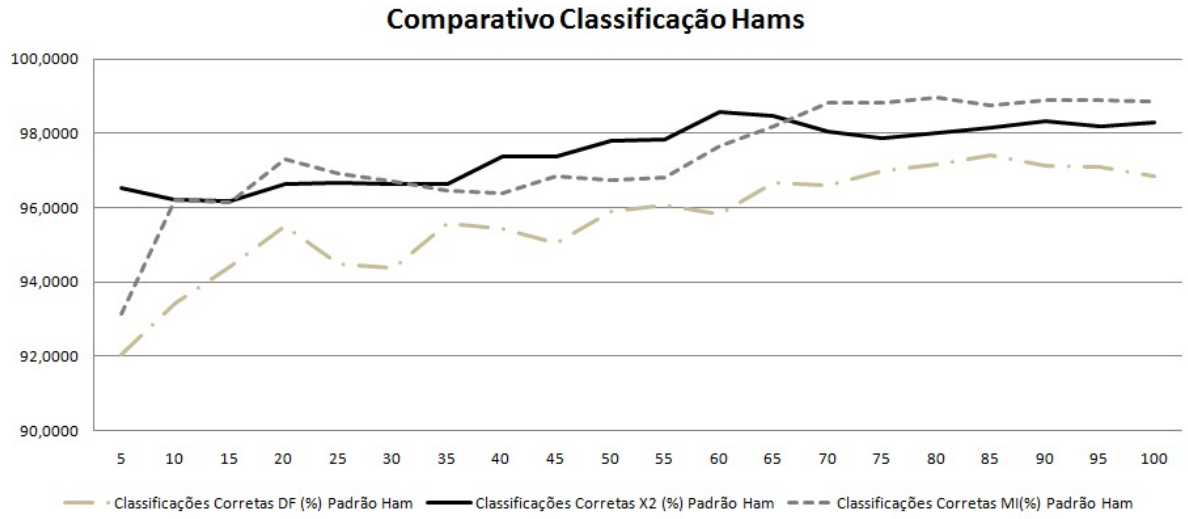
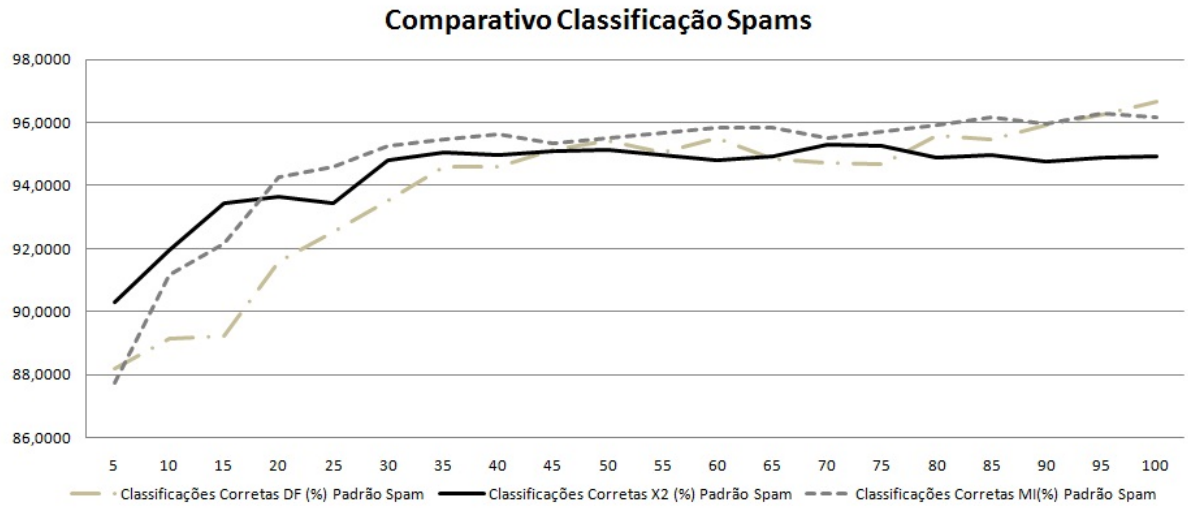


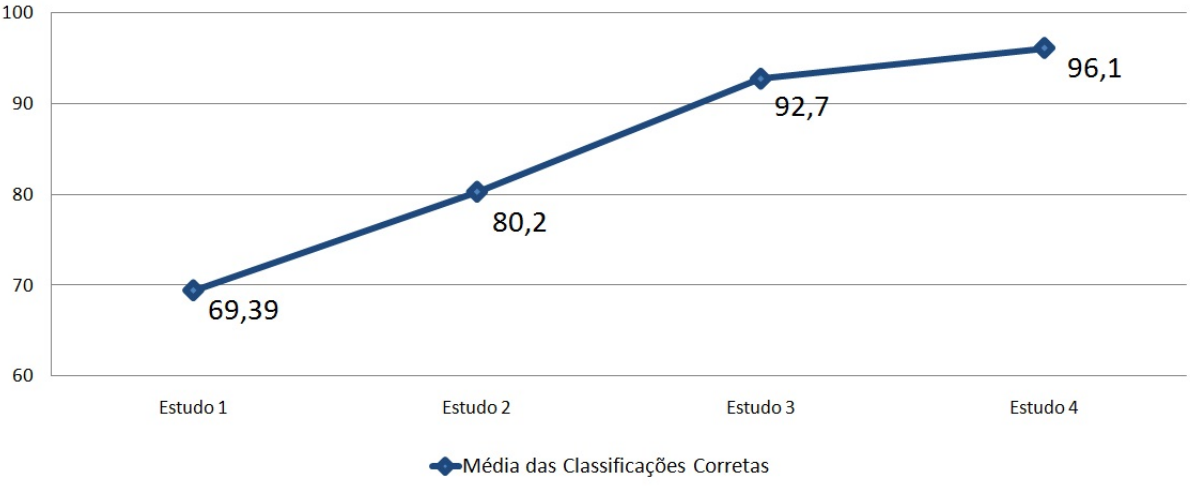
Figura 17: Comparativo Classificação Spams



Os resultados do Estudo 4 comportam-se como os do Estudo 3. A única e significante diferença consiste no fato de que todos os resultados do Estudo 4 são superiores, devido ao uso do pré-processamento.

O gráfico da figura 18 apresenta os percentuais de acerto totais obtidos nos Estudos 1 a 4. Os percentuais de acerto totais são calculados pelas médias das classificações corretas obtidas nas classificações de Spam e Ham, considerando-se todos os resultados com os diferentes tamanhos de vetores de entrada. Para os Estudos 3 e 4, o método de seleção de características considerado foi o MI, por ter apresentado o melhor desempenho médio dentre os três.

Figura 18: Evolução Classificações Corretas Estudos 1 a 4



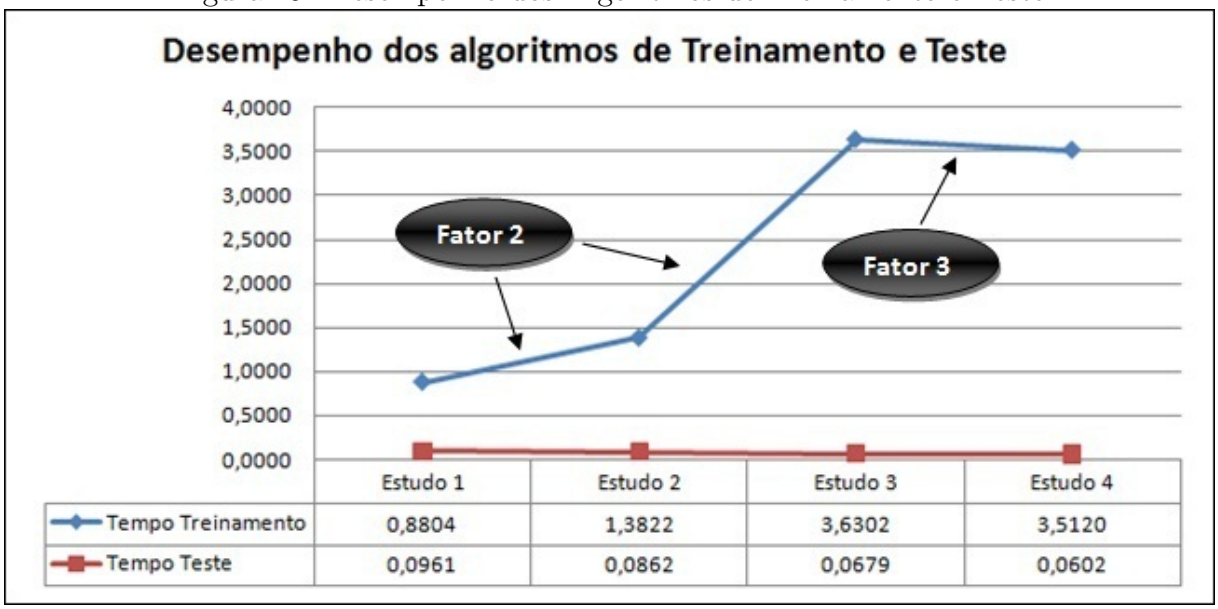
Como pode-se perceber pelo gráfico da figura 18, a mais significante melhoria no percentual de classificações corretas ocorre quando se insere os métodos de seleção de

características (Estudos 3 e 4). As técnicas de pré-processamento também produzem, porém, significativa queda no erro, quando adicionadas (Estudo 2 em relação ao Estudo 1 e Estudo 4 em relação ao Estudo 3).

Por fim, os resultados apresentados no gráfico da figura 19 consideram a média, em termos de tempos gastos no Teste, pelo classificador neural, com entradas variando de cinco a cem elementos. Analisando-se o gráfico, pode-se verificar que o tempo médio gasto no Treinamento do classificador neural é influenciado por três fatores:

- **Fator 1:** a dimensão do vetor de entrada, ou seja, o número de características selecionadas;
- **Fator 2:** o tamanho do conjunto de padrões de treinamento;
- **Fator 3:** a relevância das características selecionadas para compor o vetor de entrada, determinando uma convergência mais rápida ou não do treinamento.

Figura 19: Desempenho dos Algoritmos de Treinamento e Teste



	Estudo 1	Estudo 2	Estudo 3	Estudo 4
Tempo Treinamento	0,8804	1,3822	3,6302	3,5120
Tempo Teste	0,0961	0,0862	0,0679	0,0602

A diferença nos tempos gastos em treinamento, pelo classificador neural, entre os Estudos 1 e 2 e entre os Estudos 2 e 3, pode ser explicada pelo Fator 2. De fato, no Estudo 1, o número de padrões zerados é superior ao número de padrões zerados no Estudo 2, que, por sua vez, é superior ao número de padrões zerados no Estudo 3. Assim, o conjunto de padrões de treinamento utilizado no Estudo 1 é menor que o conjunto utilizado no Estudo 2, que, por sua vez, é bem menor que o utilizado no Estudo 3.

A diferença nos tempos gastos em treinamento, pelo classificador neural, entre os Estudos 3 e 4 pode ser explicada pelo Fator 3. O pré-processamento, ao eliminar características não significativas e ao uniformizar as demais características, como descrito na seção 2.1.3.1, reduz a quantidade total de características do universo considerado, ou seja, dos emails da Base SpamAssassin. Com menos características disponíveis, o processo de seleção de características produz características ainda mais significativas, reduzindo-se mais o número de padrões zerados gerado.

Apesar da maior redução na quantidade de padrões zerados e, por conseguinte, do aumento do tamanho do conjunto de padrões de treinamento, o tempo gasto em treinamento é inferior porque, ao utilizar características mais significativas nos padrões de treinamento, a convergência do treinamento é mais rápida.

Os tempos gastos nos testes, pelo classificador neural, apesar de bastante similares nos quatro estudos, apresentam uma tendência de queda, partindo do Estudo 1 para o Estudo 4. Este fato pode ser explicado, novamente, pela quantidade de padrões. Estudos com uma quantidade menor de padrões zerados apresentam conjuntos de teste com menor quantidade de padrões.

4.4 Estudos com a Base de Dados LingSpam

Os estudos envolvendo a Base de Dados LingSpam utilizaram tanto técnicas de pré-processamento quanto métodos de seleção de características. A opção pelo uso de ambos se deu pelo fato de que, como reportado na seção 4.3.5, as técnicas de pré-processamento aliadas aos métodos de seleção de características produzem os melhores resultados.

A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice F.

4.4.1 Estudo 5: Método DF

A tabela 28 apresenta as características relacionadas aos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 5, para o método DF.

As características mostram que a combinação do pré-processamento com o método DF, de seleção de características, produziu bons resultados na determinação das melhores características para a Base LingSpam, produzindo uma pequena quantidade de padrões zerados.

Tabela 28: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 5 - Método DF

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	963	192	0	0	0	771	1926
Validação	481	96	0	0	0	385	962
Teste	963	191	5	2	0	772	1933

A tabela 29 apresenta os resultados obtidos no Estudo 5, para o método DF, e a tabela 30, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 29: Resultados obtidos pelo Estudo 5 - DF

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	96,6943	±	0,6156	97,3875	±	0,5432
2	10	97,1443	±	0,6829	93,8438	±	1,0864
3	15	97,8634	±	0,5329	95,3389	±	1,8158
4	20	98,6342	±	0,3311	94,0817	±	0,9002
5	25	98,8774	±	0,2431	94,3197	±	1,4589
6	30	98,9188	±	0,2638	95,6286	±	1,1640
7	35	98,6756	±	0,5484	96,1873	±	1,9814
8	40	99,1102	±	0,2690	96,2028	±	0,8484
9	45	98,9860	±	0,2380	95,9390	±	1,3709
10	50	99,2550	±	0,1759	95,7217	±	1,3088
11	55	99,2757	±	0,2587	94,9250	±	1,5985
12	60	99,1205	±	0,3104	96,2390	±	1,1019
13	65	99,1671	±	0,2121	96,4977	±	0,8433
14	70	99,0998	±	0,1242	96,7563	±	1,2054
15	75	99,3533	±	0,3363	96,9477	±	1,2933
16	80	99,5189	±	0,1190	96,2183	±	0,9881
17	85	99,4982	±	0,1707	97,1029	±	0,9312
18	90	99,4258	±	0,2121	97,1443	±	0,8691
19	95	99,5034	±	0,3414	96,2338	±	0,8691
20	100	99,4775	±	0,2121	96,5235	±	1,3864

Tabela 30: Tempos em segundos gastos no Treinamento e Teste do Estudo 5 - DF

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	1,5313	0,0375
10	1,5514	0,0363
15	1,5515	0,0360
20	1,7529	0,0423
25	1,6657	0,0374
30	1,8405	0,0392
35	1,7436	0,0392
40	1,7157	0,0437
45	1,8610	0,0424
50	1,7984	0,0501
55	1,7295	0,0453
60	1,7796	0,0405
65	1,8298	0,0469
70	1,8016	0,0468
75	1,8392	0,0454
80	1,9577	0,0470
85	2,1187	0,0486
90	1,5705	0,0467
95	1,5375	0,0498
100	1,4610	0,0485

4.4.2 Estudo 5: Método Chi-Quadrado

A tabela 31 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 5, para o método Chi-Quadrado.

Tabela 31: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 5 - Método Chi-Quadrado

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	928	192	0	0	0	736	1856
Validação	464	96	0	0	0	368	928
Teste	927	193	93	0	0	734	1947

A utilização do método Chi-Quadrado não produziu padrões spam zerados. Por outro lado, a quantidade de padrões Ham zerados cresceu significativamente.

A tabela 32 apresenta os resultados obtidos no Estudo 5, para o método Chi-Quadrado, e a tabela 33, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 32: Resultados obtidos pelo Estudo 5 - Chi-Quadrado

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	98,62866	±	0,035952	91,83359	±	0,10272
2	10	94,94094	±	2,080126	96,46122	±	1,587059
3	15	98,00719	±	1,325117	94,60709	±	1,438107
4	20	98,34104	±	0,785823	94,97175	±	1,28916
5	25	98,90088	±	0,133545	95,15151	±	0,482796
6	30	98,68002	±	0,16949	95,16178	±	0,472526
7	35	98,86492	±	0,30303	94,36569	±	1,40216
8	40	99,02928	±	0,354394	94,18593	±	0,677964
9	45	98,90601	±	0,39548	93,86749	±	1,047761
10	50	98,69029	±	0,333846	94,52491	±	1,04777
11	55	98,66975	±	0,508472	94,76117	±	1,12994
12	60	98,68516	±	0,236264	95,19774	±	1,20699
13	65	98,72625	±	0,318437	95,49563	±	2,275294
14	70	98,74165	±	0,590656	95,83462	±	1,751412
15	75	98,53107	±	0,749877	95,62404	±	1,242932
16	80	98,82383	±	0,559837	95,25938	±	2,090395
17	85	98,76734	±	0,462245	96,63072	±	1,099125
18	90	98,72111	±	0,457111	96,64099	±	0,698508
19	95	98,99332	±	0,328713	96,2866	±	1,319976
20	100	99,04469	±	0,338985	96,69748	±	0,939906

Tabela 33: Tempos em segundos gastos no Treinamento e Teste do Estudo 5 - Chi-Quadrado

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	2,5109	0,0375
10	1,4550	0,0405
15	1,6188	0,0389
20	1,6985	0,0374
25	1,5216	0,0393
30	1,5423	0,0374
35	1,6217	0,0422
40	1,5703	0,0390
45	1,5657	0,0435
50	1,5392	0,0404
55	1,6001	0,0422
60	1,5860	0,0438
65	1,6016	0,0453
70	1,6203	0,0454
75	1,6111	0,0486
80	1,6360	0,0500
85	1,6561	0,0486
90	1,8030	0,0531
95	1,7501	0,0515
100	1,8110	0,0533

4.4.3 Estudo 5: Método MI

A tabela 34 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 5, para o método MI.

Tabela 34: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 5 - Método MI

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	952	192	0	0	0	760	1904
Validação	476	96	0	0	0	380	952
Teste	951	192	33	1	0	759	1936

Tal como no Estudo 4, o método MI produz uma quantidade de padrões zerados superior à produzida pelo método DF e inferior à produzida pelo método Chi-Quadrado.

A tabela 35 apresenta os resultados obtidos no Estudo 5, para o método MI, e a tabela 36, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 35: Resultados obtidos pelo Estudo 5 - MI

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	5	88,0217	±	1,9163	96,2293	±	3,3058
2	10	97,2417	±	0,4339	92,0558	±	0,6818
3	15	97,5155	±	0,7800	92,4897	±	0,8988
4	20	97,8048	±	0,3357	91,4773	±	0,9297
5	25	97,9855	±	0,8264	94,2562	±	0,8884
6	30	98,1353	±	0,5217	94,0857	±	1,5754
7	35	98,7448	±	0,3254	93,8843	±	0,9607
8	40	98,8481	±	0,3771	94,2149	±	1,2913
9	45	98,7552	±	0,5733	94,6746	±	1,8027
10	50	98,7242	±	0,5837	94,8760	±	1,3017
11	55	99,0754	±	0,3048	93,7448	±	1,3895
12	60	98,4246	±	0,7490	95,2376	±	1,1467
13	65	98,8378	±	0,3874	95,0155	±	1,4205
14	70	98,6519	±	0,7283	95,9659	±	1,9680
15	75	98,7810	±	0,4855	95,9659	±	1,4515
16	80	98,5434	±	0,5269	96,5238	±	0,8110
17	85	99,0702	±	0,4649	95,7800	±	1,3585
18	90	98,9618	±	0,4700	96,1829	±	1,2862
19	95	98,8791	±	0,5010	96,5186	±	1,2087
20	100	98,6467	±	0,8368	96,6322	±	1,0950

Tabela 36: Tempos em segundos gastos no Treinamento e Teste do Estudo 5 - MI

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
5	1,8682	0,0481
10	1,9650	0,0482
15	1,8251	0,0407
20	1,8202	0,0532
25	1,8700	0,0560
30	1,5686	0,0391
35	1,6795	0,0407
40	1,7139	0,0407
45	1,6282	0,0423
50	1,6688	0,0424
55	1,9047	0,0467
60	1,6593	0,0469
65	1,7766	0,0452
70	1,6641	0,0483
75	1,7310	0,0454
80	1,7045	0,0469
85	1,8282	0,0468
90	1,5811	0,0469
95	1,6642	0,0500
100	1,6622	0,0485

4.4.4 Análise dos resultados sobre a Base LingSpam (Estudo 5)

Os gráficos das figuras 20 e 21 apresentam os percentuais de classificações corretas de Spam e Ham obtidas pelos três métodos no Estudo 5, variando com o número de características selecionadas.

Analisando os gráficos, pode-se verificar que, tanto na classificação de emails Ham quanto Spam, o erro médio decresce com o aumento da dimensionalidade do vetor de entrada. Na classificação de emails Ham, porém, o erro médio tende a estabilização após um determinado valor de dimensionalidade.

Na classificação de Hams, os três métodos apresentaram desempenhos similares. Para os padrões de spam, porém, é perceptível visualizar-se, no gráfico, um desempenho melhor do método DF. Tal fato se dá, muito provavelmente, pelo fato da Base LingSpam originar-se de poucas fontes de informação, acentuando a distância entre as classes Spam e Ham. Portanto, tal característica, favorece a classificação pelo método DF, uma vez que este não leva em conta a relevância de uma característica para com as duas classes (Spam e Ham).

Figura 20: Comparativo Classificação Hams

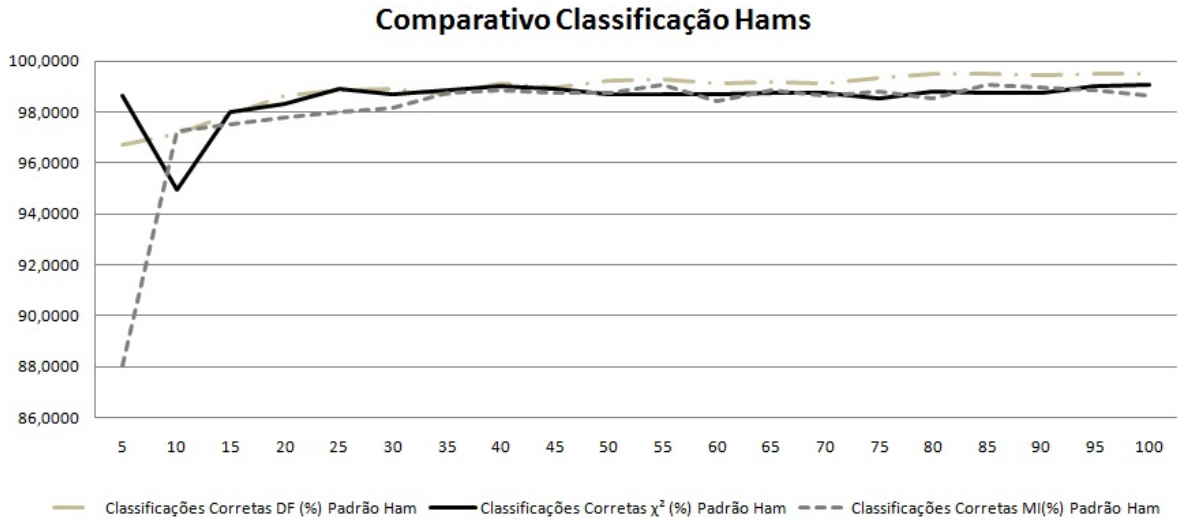
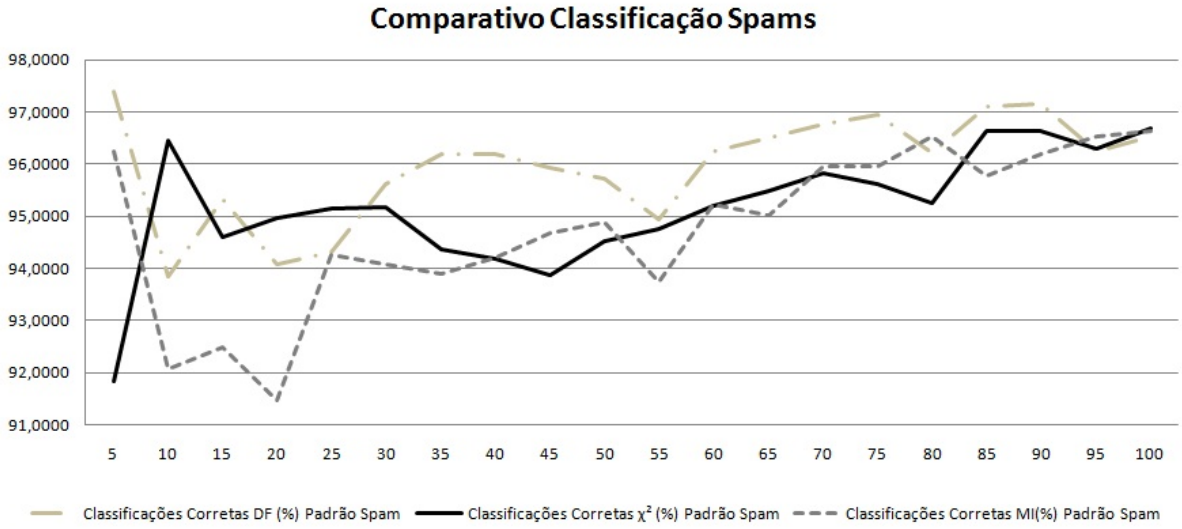


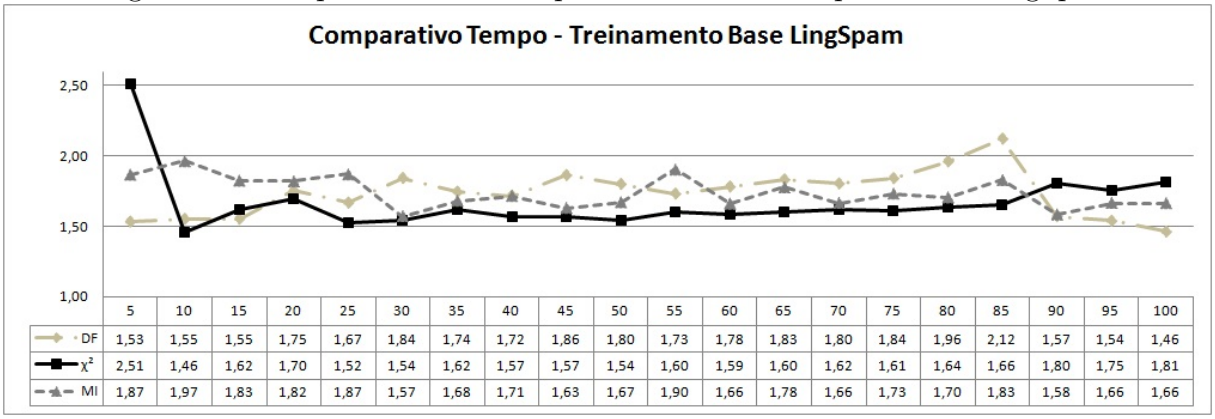
Figura 21: Comparativo Classificação Spams



De modo geral, os resultados obtidos com a Base LingSpam mostram-se superiores aos obtidos com a Base SpamAssassin. Isto pode ser explicado pelo fato da Base LingSpam ser menor e mais uniforme quanto as fontes de informação, facilitando o trabalho de classificação do classificador neural.

O gráfico da figura 22 apresenta a comparação nos tempos do classificador neural para os três métodos abordados. O gráfico mostra que o menor tempo médio de treinamento foi obtido pelo método Chi-Quadrado. Entretanto, este foi o de pior desempenho nas extremidades, ou seja, foi superado pelos métodos MI e DF tanto para vetores de entrada com 5 características quanto para vetores de entrada com mais de 85 características.

Figura 22: Comparativo dos Tempos de Treinamento para Base LingSpam



	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
DF	1,53	1,55	1,55	1,75	1,67	1,84	1,74	1,72	1,86	1,80	1,73	1,78	1,83	1,80	1,84	1,96	2,12	1,57	1,54	1,46
χ^2	2,51	1,46	1,62	1,70	1,52	1,54	1,62	1,57	1,57	1,54	1,60	1,59	1,60	1,62	1,61	1,64	1,66	1,80	1,75	1,81
MI	1,87	1,97	1,83	1,82	1,87	1,57	1,68	1,71	1,63	1,67	1,90	1,66	1,78	1,66	1,73	1,70	1,83	1,58	1,66	1,66

Os tempos gastos nos testes, pelo classificador neural, apresentam desempenho dentro do esperado. O tempo gasto cresce com o número de características presentes no vetor de entrada.

4.5 Estudo 6: Testes utilizando a Base de Dados Trec

Os estudos envolvendo a Base de Dados Trec utilizaram, igualmente, tanto técnicas de pré-processamento quanto métodos de seleção de características. A Base Trec (TREC, 2007) originou-se do Spam Track, um evento da conferência anual cujo escopo é a pesquisa sobre recuperação de texto (TREC – Text Retrieval Conference).

A Base de Dados Trec contém uma extensa coletânea de emails, incluindo emails com imagens. Por este motivo já foi empregada em trabalhos envolvendo classificação de emails com imagens (BYUN et al., 2007). Nenhum tratamento específico para as imagens, porém, foi realizado, mantendo-se as mesmas técnicas de pré-processamento empregadas nos estudos anteriores.

Devido ao grande número de emails e, consequentemente, de padrões presentes na Base, os tempos computacionais gastos nos treinamentos e testes do classificador neural foram longos. Devido a isto, os testes com a Base de Dados Trec iniciaram com vetores de dez elementos, variando de dez em dez elementos até atingir cem elementos.

A lista completa com todas a palavras consideradas no estudo é apresentada no apêndice G.

4.5.1 Estudo 6: Método DF

A tabela 37 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 6, para o método DF.

Tabela 37: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 6 - Método DF

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	10.021	19.841	-	-	9.820	-	39.682
Validação	5.011	9.920	-	-	4.909	-	19.840
Teste	10.021	19.841	161	472	9.820	-	40.315

Apesar das características mostrarem uma quantidade elevada de padrões zerados, em termos percentuais, sua quantidade é inferior a 1% do total de emails da Base.

A tabela 38 apresenta os resultados obtidos no Estudo 6, para o método DF, e a tabela 39, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 38: Resultados obtidos pelo Estudo 6 - DF

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	10	94,07937	±	1,43595	90,88032	±	0,69553
2	20	96,29493	±	0,66154	92,03522	±	0,90785
3	30	97,85464	±	0,22597	93,53839	±	1,08893
4	40	97,87350	±	0,22894	93,90277	±	0,65757
5	50	97,82959	±	0,18852	94,92174	±	0,54645
6	60	98,02282	±	0,39663	95,64232	±	0,83790
7	70	97,99578	±	0,36959	95,38633	±	0,47129
8	80	98,17041	±	0,26144	96,26169	±	1,12142
9	90	98,34454	±	0,27186	96,05110	±	0,87560
10	100	98,20861	±	0,16322	96,65162	±	0,53305

Tabela 39: Tempos em segundos gastos no Treinamento e Teste do Estudo 6 - DF

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
10	147,7795	0,4281
20	168,5093	0,4564
30	174,0204	0,4828
40	172,3940	0,5311
50	177,0562	0,5547
60	152,2687	0,6080
70	144,4610	0,6359
80	143,6905	0,6642
90	142,7562	0,6936
100	130,9468	0,7437

4.5.2 Estudo 6: Método Chi-Quadrado

A tabela 40 apresenta as características dos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 6, para o método Chi-Quadrado.

Tabela 40: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 6 - Método Chi-Quadrado

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	10.036	19.692	-	-	9.656	-	39.384
Validação	5.018	9.846	-	-	4.828	-	19.692
Teste	10.037	19.692	123	844	9.655	-	40.351

O método Chi-Quadrado apresentou uma taxa inferior de padrões Ham zerados em relação ao método DF, porém apresentou uma quantidade bem superior de padrões Spam zerados.

A tabela 41 apresenta os resultados obtidos no Estudo 6, para o método Chi-Quadrado, e a tabela 42, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 41: Resultados obtidos pelo Estudo 6 - Chi-Quadrado

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	10	90,5251	±	0,6562	95,4933	±	0,3061
2	20	96,4011	±	0,3400	96,5785	±	1,0047
3	30	97,5887	±	0,1537	98,5078	±	0,2233
4	40	97,9725	±	0,1980	98,5341	±	0,3556
5	50	98,0610	±	0,2067	98,6186	±	0,4025
6	60	98,0843	±	0,1363	98,5071	±	0,3182
7	70	98,2855	±	0,1482	98,5911	±	0,3383
8	80	97,7245	±	0,7018	98,3822	±	0,2969
9	90	97,3247	±	1,2106	98,4464	±	0,3715
10	100	97,4184	±	0,7864	98,5195	±	0,4624

Tabela 42: Tempos em segundos gastos no Treinamento e Teste do Estudo 6 - Chi-Quadrado

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
10	208,0830	0,4203
20	251,7717	0,4500
30	180,4828	0,4844
40	153,8186	0,5206
50	128,6250	0,5483
60	125,2406	0,5861
70	116,2733	0,6202
80	99,4829	0,6501
90	87,6266	0,6783
100	94,7984	0,7252

4.5.3 Estudo 6: Método MI

A tabela 43 apresenta as características relacionadas aos conjuntos de dados de Treinamento, Validação e Teste utilizadas no Estudo 6, para o método MI.

As características indicam que a quantidade de padrões Ham zerados é superior à produzida pelos métodos anteriores. No entanto, a quantidade de padrões Spam zerados situa-se entre a quantidade produzida pelo método DF e a produzida pelo método Chi-Quadrado.

A tabela 44 apresenta os resultados obtidos no Estudo 6, para o método MI, e a tabela 45, os tempos gastos no Treinamento e Teste do classificador neural.

Tabela 43: Padrões nos conjuntos de Treinamento, Validação e Teste do Estudo 6 - Método MI

	Padrões		Padrões Zerados		Padrões Duplicados		Total
	Ham	Spam	Ham	Spam	Ham	Spam	
Treinamento	9.995	19.769	-	-	9.774	-	39.538
Validação	4.997	9.884	-	-	4.887	-	19.768
Teste	9.995	19.770	227	651	9.775	-	40.418

Tabela 44: Resultados obtidos pelo Estudo 6 - MI

Experimento	No. Entradas	Classificações Corretas (%)					
		Padrão *Ham*			Padrão *Spam*		
1	10	93,2157	±	1,4592	83,6200	±	1,1376
2	20	94,9166	±	0,4743	93,9703	±	0,4532
3	30	95,8313	±	0,4874	94,7133	±	0,4555
4	40	96,6841	±	0,5376	95,4545	±	0,5821
5	50	96,9292	±	0,1267	96,0643	±	0,4745
6	60	97,2234	±	0,1848	96,5443	±	0,6700
7	70	97,0332	±	0,3345	97,1675	±	0,5341
8	80	97,2835	±	0,8511	97,3513	±	0,9144
9	90	97,8112	±	0,2664	97,8525	±	0,7199
10	100	97,8808	±	0,3528	98,2284	±	0,3788

Tabela 45: Tempos em segundos gastos no Treinamento e Teste do Estudo 6 - MI

Nº de Entradas	Tempo Médio Treinamento	Tempo Médio Teste
10	163,2905	0,4331
20	223,7292	0,4643
30	174,8027	0,4954
40	162,5479	0,5485
50	169,8003	0,5892
60	184,3337	0,6253
70	149,0521	0,6601
80	129,5294	0,6962
90	139,9709	0,7290
100	134,8670	0,7680

4.5.4 Análise dos resultados sobre a Base Trec (Estudo 6)

Os gráficos das figuras 23 e 24 apresentam os percentuais de classificações corretas de Spam e Ham obtidas pelos três métodos no Estudo 6, variando com o número de características selecionadas.

Figura 23: Comparativo Classificação Hams

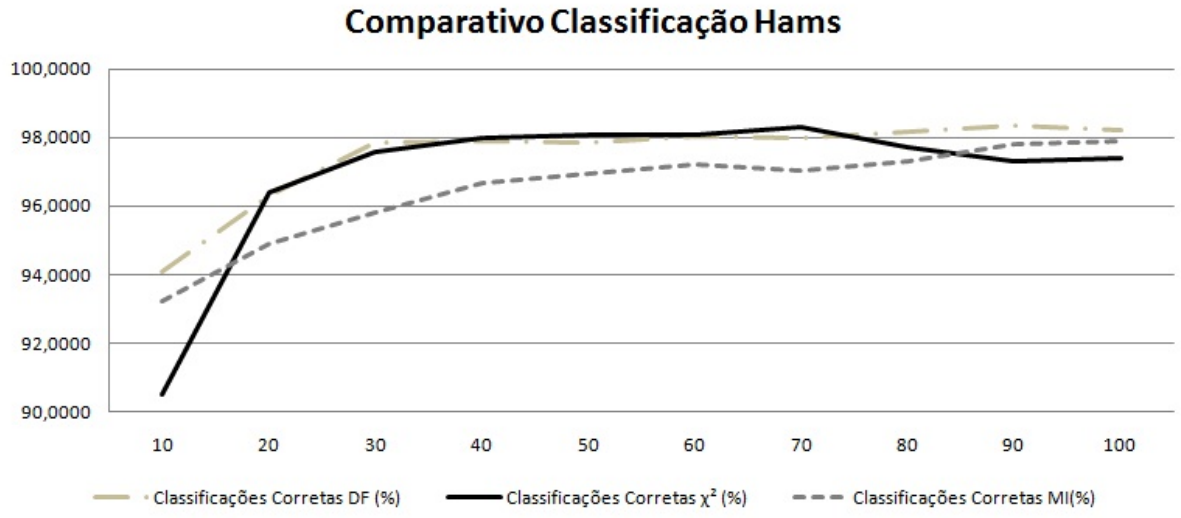
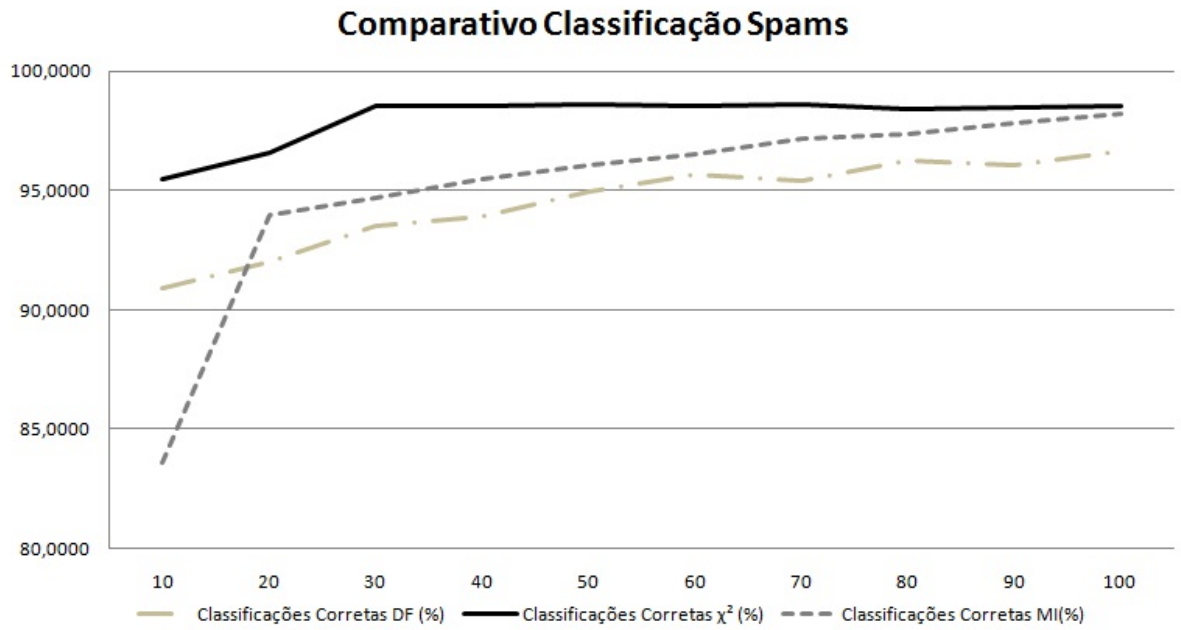


Figura 24: Comparativo Classificação Spams



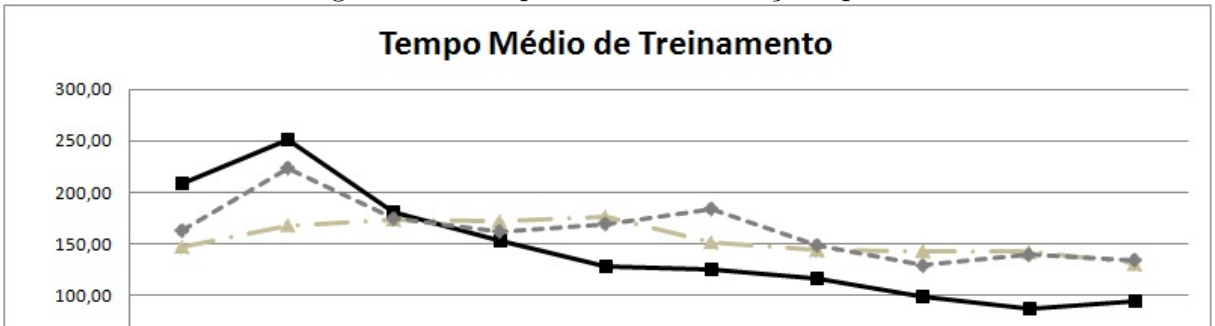
Como ocorrido com a Base de Dados SpamAssassin, o erro médio tende a estabilizar-se a partir de um determinado valor de dimensionalidade do vetor de entrada.

Os gráficos indicam, igualmente, que não houve um método superior em todas as situações. Na classificação de Hams, o método de melhor desempenho, em média, foi o DF, sendo superado, porém, pelo método Chi-Quadrado em algumas poucas dimensões de vetores de entrada.

Para a classificação de emails Spam, o método de melhor desempenho foi, indiscutivelmente, o Chi-Quadrado, que, dado o bom desempenho também na classificação de emails Ham, pode ser considerado aquele que alcançou o melhor desempenho global sobre a Base de Dados Trec. Por outro lado, o método MI, que apresentou desempenho satisfatório para as Bases SpamAssassin e LingSpam, não obteve o mesmo desempenho para a Base de Dados Trec.

Por fim, como observado nos gráficos e nas tabelas de resultados do Estudo 6, ainda que a Base Trec apresente uma quantidade maior de emails com formatos diversos (incluindo imagens), o classificador neural foi capaz de identificar satisfatoriamente a ocorrência de emails Spam e Ham, com percentuais de classificação corretas acima de 98%.

Figura 25: Comparativo Classificação Spams



	10	20	30	40	50	60	70	80	90	100
DF	147,78	168,51	174,02	172,39	177,06	152,27	144,46	143,69	142,76	130,95
χ^2	208,08	251,77	180,48	153,82	128,63	125,24	116,27	99,48	87,63	94,80
MI	163,29	223,73	174,80	162,55	169,80	184,33	149,05	129,53	139,97	134,87

O gráfico da figura 25 mostra os tempos gastos no treinamento do classificador neural para os três métodos de seleção de características utilizados. Como pode-se perceber, a partir de trinta características no vetor de entrada, o método Chi-Quadrado apresenta tempos de treinamento inferiores aos dos outros métodos. Com melhor seleção de características, o treinamento do classificador converge mais rapidamente para erros menores, o que, por sua vez, produz resultados mais acurados nos testes.

Outro ponto importante a ser observado é que a partir de 60 características presentes no vetor de entrada, todos os métodos apresentam queda nos tempos de treinamento, sugerindo que as características agregadas a partir deste ponto são também relevantes

para a classificação.

Os tempos gastos nos testes do classificador neural apresentam o comportamento esperado. Os tempos crescem com o aumento da dimensionalidade do vetor de entrada.

4.6 Comparativo entre os Métodos de Seleção de Características

Como apresentado pelos Estudos de 1 a 6 nenhum dos três métodos de seleção de características empregados foi o melhor para todas as situações. Tal fato decorre principalmente das peculiaridades das fontes de informação utilizadas para formação das Bases de Dados utilizadas.

A tabela 46 apresenta o desempenho médio dos métodos empregados sobre as três bases de dados utilizadas. O desempenho médio é mensurado através da média de classificações corretas para todas as variações de tamanho do vetor de entrada, obtidas pelo método para uma determinada classe (Spam e Ham) sobre a base de dados avaliada.

Tabela 46: Comparativo dos Métodos de Seleção de Características

	SpamAssassin			LingSpam			Trec		
Métodos	Ham	Spam	Precisão	Ham	Spam	Precisão	Ham	Spam	Precisão
DF	95,70	93,93	94,99	98,88	95,96	**97,71**	97,47	94,53	96,29
χ^2	97,49	94,37	96,24	98,53	95,18	97,19	96,94	98,02	**97,37**
MI	97,38	94,81	**96,35**	97,98	94,79	96,71	96,48	95,10	95,93

A tabela 46 apresenta, igualmente um indicador de precisão, calculado a partir da equação 4.1.

$$Precisao = 0,6 * (ClassificacoesCorretasHam) + 0,4 * (ClassificacoesCorretasSpam) \tag{4.1}$$

Para o cálculo da precisão, são utilizadas ponderações maiores para classificações corretas de emails Ham devido ao problema dos falsos positivos, ou seja, classificações incorretas de emails Ham são tidas como mais custosas que classificações incorretas de emails Spam.

A partir da tabela 46 pode-se observar que apesar de o método MI ter obtido o melhor desempenho para a Base SpamAssassin, o mesmo obteve desempenho inferior aos outros

dois métodos para as Bases de Dados LingSpam e Trec, onde os métodos de melhor desempenho foram DF e Chi-Quadrado respectivamente.

O fato de não existir, entre os casos estudados, um método ideal, com melhor desempenho para todas as situações, justifica novas pesquisas envolvendo outros métodos para seleção de características em emails.

4.7 Estudo 7: Teste de obfuscação

A seção 2.1.4.2 descreve um série de técnicas utilizadas por *spammers* na tentativa de "driblar" os filtros antispam. O sistema antispam desenvolvido não implementa procedimentos para tratar a maioria destas técnicas de obfuscação, tais como o uso de tags HTML inválidas, uso de texto invisível, uso de tags HTML em branco, uso de comentários HTML, dentre outras. O desenvolvimento de procedimentos para tratamento destas técnicas de obfuscação é importante e será abordado em trabalhos futuros.

O sistema antispam desenvolvido implementa, porém, procedimentos para tratar uma das técnicas mencionadas na seção 2.1.4.2, qual seja, a técnica que visa esconder palavras com caracteres inválidos. O estudo 7 foi realizado com o intuito de verificar-se o desempenho do sistema antispam na presença de emails que façam uso destas técnicas de obfuscação.

No estudo 7, utilizou-se o classificador neural desenvolvido no estudo 4, ou seja, treinado com o mesmo procedimento e dados deste estudo. Em seguida, dez emails ham foram escolhidos aleatoriamente. Em cada um destes dez emails foi inserida uma palavra com caracteres inválidos, tais como "V I A G R A", "V*I@GR@", dentre outras.

Os dez emails foram submetidos ao sistema antispam para verificar se o mesmo classificaria-os, desta vez, como emails spam. Para fins estatísticos, para cada dimensão do vetor de entrada, o processo de treinamento do classificador neural foi realizado por dez vezes. O apêndice H traz a lista completa das 100 palavras escolhidas pelo método MI para esse estudo.

A tabela 47 mostra o percentual de classificações corretas, ou seja, como spams, destes dez emails.

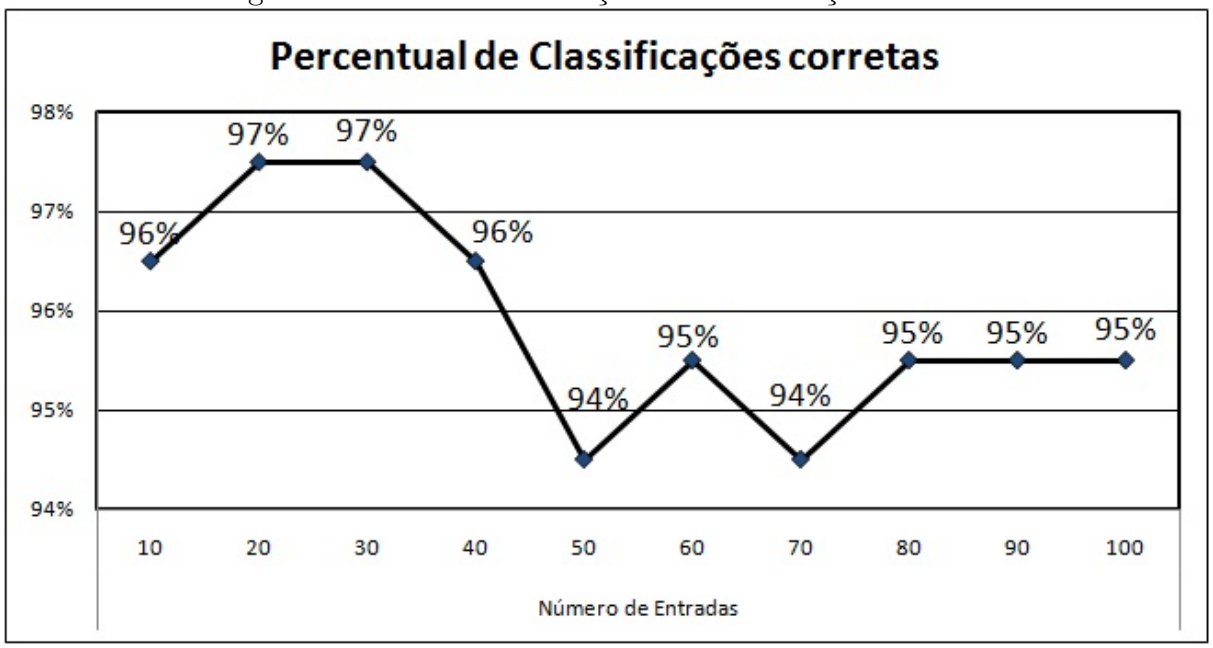
Pelos resultados, pode-se verificar que a grande maioria destes dez emails foi classificada corretamente como spam. À medida em que se aumenta a dimensionalidade do vetor de entrada, o percentual de classificações corretas apresenta uma leve redução. Isto

Tabela 47: Resultado Classificação Estudo 7

	Número de Entradas									
	10	20	30	40	50	60	70	80	90	100
Percentual de Classificações corretas	96%	97%	97%	96%	94%	95%	94%	95%	95%	95%

é devido ao fato de se utilizarem, nos vetores de entrada, mais características relativas a emails ham, uma vez que estes dez emails eram, originariamente, emails ham. O gráfico 26 apresenta a variação do percentual da classificação segundo a dimensionalidade do vetor de entrada.

Figura 26: Estudo 7: Evolução das classificações corretas



5 Conclusão

O crescimento da Internet e sua popularização tornaram-na alvo de ataques. Dentre estes, podem-se citar os spams, emails não-solicitados enviados em larga escala pela rede.

Emails spam, disseminados por indivíduos conhecidos como *spammers*, são hoje um sério problema, posto que, afetam tanto o funcionamento da Internet quanto seus usuários. O volume de emails spam impacta diretamente em caixas de correio preenchidas com mensagens indesejadas, perdas de eficiência, gastos com o tráfego e tempo para eliminação dos emails indesejados, o que, por sua vez, acarreta sérios prejuízos financeiros às corporações e aos próprios usuários.

Este trabalho propõe um sistema antispam composto por três estágios: pré-processamento, seleção de características e classificação. No estágio de pré-processamento, os emails são uniformizados, conteúdos irrelevantes são removidos e alguns padrões de spam conhecidos são postos em relevo. No estágio de seleção de características, as características dos emails são ordenadas segundo sua relevância. No estágio de classificação, realizado por intermédio de uma rede neural artificial, os emails são classificados em Ham ou Spam.

O sistema proposto foi exaustivamente testado, sobre três bases públicas de dados, em uma série de estudos. Os estudos empregaram três métodos estatísticos de seleção de características – distribuição de frequência, chi-quadrado e informação mútua –, amplamente utilizados em classificação de textos, na área de aprendizagem de máquina (machine learning). Como rede neural artificial, foi utilizado um perceptron multicamadas, por sua simplicidade, rapidez no treinamento e reconhecida capacidade de generalização.

Os estudos visaram, igualmente, mensurar a contribuição de cada um dos três estágios no resultado final de classificação dos emails. Desta forma, pelos resultados obtidos, comprovou-se que tanto o estágio de pré-processamento quanto o de seleção de características são relevantes para a precisão nas classificações.

Dentre os três métodos estatísticos de seleção de características empregados, o método MI foi o que apresentou melhor desempenho para a Base SpamAssassin. Foi superado,

porém, pelo método DF na classificação dos emails da Base de Dados LingSpam, e pelos métodos Chi-Quadrado e DF na classificação dos emails da Base de Dados Trec, conforme relatado na seção 4.6. Este fato mostra que nenhum dos métodos empregados é o melhor para todas as situações, abrindo assim, um novo campo para pesquisas futuras.

O classificador neural foi empregado com vinte diferentes quantidades de neurônios na camada de entrada, para as Bases de Dados SpamAssassin e LingSpam, e com dez diferentes quantidades de neurônios, para a Base de Dados Trec, de acordo com a dimensionalidade do vetor de entrada utilizado. Apesar de intuitivo o fato de que maiores dimensionalidades do vetor de entrada elevem a precisão das classificações, verificou-se que, a partir de um certo valor de dimensionalidade, o ganho obtido tende a ser muito pequeno.

Outra contribuição importante do trabalho foi a de mensurar os tempos de treinamento e teste do classificador neural. Uma vez mais, apesar de intuitivo o fato de que maiores dimensionalidades do vetor de entrada exijam maiores tempos de treinamento e teste, verificou-se que a utilização de vetores de entrada com maiores dimensionalidades, contendo, porém, as características mais relevantes selecionadas pelos métodos de seleção de características, melhoram a convergência e a qualidade do treinamento, gerando menores tempos de treinamento.

Os resultados obtidos pelo sistema antispam sobre as três bases de dados são promissores. Os resultados são superiores aos reportados, sobre as mesmas bases de dados, na literatura. Os tempos de processamento exigidos para o pré-processamento, seleção de características e treinamento do classificador neural são muito pouco significantes, o que permite a transformação do sistema antispam desenvolvido em uma ferramenta para uso nas corporações.

Por fim, é importante mencionar que o sistema antispam foi todo desenvolvido sobre uma arquitetura modularizada e orientada a objetos. Assim, sua arquitetura facilita a adição de novos componentes como, por exemplo, para pré-processamento de novos padrões criados por *spammers*. Este fato é de fundamental importância, uma vez que sistemas antispam devem ser capazes de evoluir em suas técnicas de classificação, em consonância com a evolução das técnicas utilizadas pelos *spammers*.

5.1 Trabalhos Futuros

Como sugestões para trabalhos futuros são elencados os seguintes itens:

- Estudo de novos métodos estatísticos de seleção de características;
- Estudo das características dos emails, classificando-as em categorias, de forma a utilizarem-se métodos estatísticos de seleção de características mais apropriados para cada uma destas categorias;
- Desenvolvimento de um módulo, para o sistema antispam, com a finalidade de identificar emails spam em imagens;
- Desenvolvimento de um módulo, para o sistema antispam, com a finalidade de apresentar ao sistema antispam exatamente o email apresentado ao usuário. Isto permite a identificação de padrões, utilizados por *spammers*, que, como descrito em (COURNANE; HUNT, 2004), são ignorados pelos sistemas antispam atuais;
- Desenvolvimento, para o estágio classificador, de uma rede neural (perceptron multicamadas) ou de um modelo baseado em kernel (support vector machine) na linguagem Java, de forma a que o sistema antispam fique todo codificado nesta linguagem;
- Por último, a transformação do sistema antispam desenvolvido em uma ferramenta tecnológica para uso nas corporações.

Referências

ANDROUTSOPOULOS, I. et al. An evaluation of naive bayesian anti-spam filtering. 2000.

ANDROUTSOPOULOS, I. et al. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. 2000.

ANTISPAM, C. G. da Internet no B. Acessado em 15/02/2011. Página na Internet - http://www.antispam.br.

BYUN, B. et al. A discriminative classifier learning approach to image modeling and spam image identification. *Fourth Conference on Email and Anti-Spam - CEAS*, 2007.

CGI, C. G. da Internet no B. Acessado em 15/02/2011. Página na Internet - http://www.cgi.br.

CHEN, S. Y.; WANG, C. C. Using header session messages to anti-spamming. 2007.

CHUAN, Z. et al. A lvq-based neural network anti-spam approach. 2005.

COURNANE, A.; HUNT, R. An analysis of the tools used for the generation and prevention of spam. *Computer e Security (2004) 23, 154 a 166*, 2004.

DESHPANDE, A.; PARK, J. S. Spam detection: Increasing accuracy with a hybrid solution. 2006.

ELKAN, C. Naive bayesian learning. *Adapted from Technical Report No. CS97-557, Department of Computer Science and Engineering, San Diego: University of California*, 1997.

FRANK, E. et al. Using model trees for classification. *Machine Learning, 63 a 76*, 1998.

FREUND, Y.; MASON, L. The alternating decision tree learning algorithm proceeding. *Sixteenth international conference on machine learning, Bled, Slovenia, 124 a 133*, 1999.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation.* 2ª. ed.. ed. [S.l.]: Prentice-Hall., 1999.

JESSEN, K. S.; CHAVES, M. H. P. C.; HOEPERS, C. Projeto de desenvolvimento de um sistema de controle e acompanhamento e notificações de spam. *V Simpósio de Segurança em Informática - São José dos Campos*, 2003.

KIM, B. M.; KANG, S. J.; KIM, J. W. A fuzzy inference method for spam-mail filtering. 2005.

LEE, P. Y.; HUI, S. C.; FONG, A. M. Neural networks for web content filtering. *IEEE Intelligent Systems (2002) 48-57*, 2002.

LYNAM, T.; CORMACK, G. Spam corpus creation for trec. 2005.

MALCON, A.; XEREZ, M. *Redes Neurais Artificiais Introdução e Princípios de Neurocomputação*. [S.l.]: Editora Eko, 1996.

OZGUR, L.; GUNGOR, T.; GURGEN, F. Adaptive anti-spam filtering for agglutinative languages: a special case for turkish. *Pattern Recognition Letters 25, p. 1819-1831.*, 2004.

PAPOULIS, A.; PILLAI, S. U. *Probability, Random Variables, and Stochastic Processes.* 4. ed. [S.l.]: McGraw-Hill, 2001.

RUMELHART, D. E.; MCCLELLAND, J. L. *Parallel Distributed Processing.* [S.l.]: The MIT Press., 1986.

RUSSEL, S. J.; NORVIG, P. *Artificial Intellligence: a modern approach.* [S.l.]: Prentice-Hall., 1995.

SAKKIS, G. et al. A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval, v. 6, n. 1. (2003) 49-73.*, 2003.

SILVA, A. M.; MOITA, G. F.; ALMEIDA, P. E. M. Um filtro anti-spam utilizando redes neurais artificiais multilayer perceptron. *V Simpósio de Segurança em Informática - São José dos Campos*, 2003.

SPAMASSASSIN. *Spamassassin Project*. 2008. Página na Internet, Disponível em: <http://spamassassin.apache.org/>.

TEMPLETON, B. *Origin of the term "spam"to mean net abuse*. Acessado em 22/01/2011. Página na Internet - http://www.templetons.com/brad/spamterm.html.

TREC. *TREC Public Spam Corpus*. 2007. Página na Internet, Disponível em: http://plg.uwaterloo.ca/ gvcormac/treccorpus07/.

TRESP, V. *Committee machines*. [S.l.]: USA: CRC Press, 2001.

VAPNIK, V. N.; WU, D.; DRUCKER, H. Support vector machines for spam categorization. 1999.

WEB, A. *Preferência do e-mail na comunicação corporativa*. 2003. Página na Internet - - http://www.agenciaweb.com.br/noticias.asp?pIdMateria=622.

XU, H.; YU, B. Automatic thesaurus construction for spam filtering using revised back propagation neural network. 2010.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. *In Proceedings of the International Conference on Machine Learning.*, 1997.

ZORKADIS, V.; KARRAS, D. A.; PANAYOTOU, M. Efficient information theoretic strategies for classifier combination, feature extration and performance evaluation in improving false positives and false negatives for spam e-mail filtering. 2005.

APÊNDICE A – Primeiro Spam Publicado

<Segue a cópia da mensagem enviada>

Newsgroups: fr.comp.os.linux

Message-ID: <1785@lcanter.win.net>

Reply-To: cslaw@lcanter.win.net (Laurence A Canter)

From: cslaw@lcanter.win.net (Laurence A Canter)

Date: Sat, 05 Mar 1994 22:48:40 GMT

Subject: U.S. Green Card Lottery - New Immigration Opportunity

1994 Green Card Lottery Information and Assistance available. The Green Card Lottery is a completely legal program giving away Green Cards to persons born in certain countries. PERSONS BORN IN MOST COUNTRIES QUALIFY, MANY FOR FIRST TIME. Only countries not qualifying are: Mexico; India; P.R. China; Philippines & Korea. There are also several countries that are uncertain at this time. They are: Canada; U.K.; Taiwan; Jamaica; Viet Nam and Dominican Republic. Lottery registration will take place soon. 55,000 Green Cards will be given to those who register correctly. NO JOB IS REQUIRED. THERE IS A STRICT DEADLINE. THE TIME TO START IS NOW!!

For FREE information via Email, send request to

Canter & Siegel, Immigration Attorneys

E Camelback Road, Ste 250, Phoenix AZ 85018 USA

cslaw@indirect.com telephone (602)661-3911 Fax (602) 451-7617

From: nike@indirect.com (Laurence Canter)

Newsgroups: comp.graphics.animation,fr.comp.os.linux

Subject: Green Card Lottery- Final One?

Date: 12 Apr 1994 07:52:28 GMT

Organization: Canter & Siegel

Message-ID: <2odjvs$2ur@herald.indirect.com>

NNTP-Posting-Host: id1.indirect.com

Green Card Lottery 1994 May Be The Last One! THE DEADLINE HAS BEEN ANNOUNCED. The Green Card Lottery is a completely legal program giving away a certain annual allotment of Green Cards to persons born in certain countries. The lottery program was scheduled to continue on a permanent basis. However, recently, Senator Alan J Simpson introduced a bill into the U. S. Congress which could end any future lotteries. THE 1994 LOTTERY IS SCHEDULED TO TAKE PLACE SOON, BUT IT MAY BE THE VERY LAST ONE. PERSONS BORN IN MOST COUNTRIES QUALIFY, MANY FOR FIRST TIME. The only countries NOT qualifying are: Mexico; India; P.R. China; Taiwan, Philippines, North Korea, Canada, United Kingdom (except Northern Ireland), Jamaica, Domican Republic, El Salvador and Vietnam. Lottery registration will take place soon. 55,000 Green Cards will be given to those who register correctly. NO JOB IS REQUIRED. HERE IS A STRICT JUNE DEADLINE. THE TIME TO START IS NOW!!

For FREE information via Email, send request to cslaw@indirect.com

Canter & Siegel, Immigration Attorneys E Camelback Road, Ste 250, Phoenix AZ 85018 USA cslaw@indirect.com telephone (602)661-3911 Fax (602) 451-7617

<Segue a cópia da resposta padrão enviada por eles>

From witch!lcanter!cslaw Thu Mar 17 21:50:25 1994 Return-Path: <witch!lcanter!cslaw> Received: from witch by nataa.frmug.fr.net with uucp (Linux Smail3.1.28.1 #14) id m0phP1N-000KmdC; Thu, 17 Mar 94 21:50 GMT+0100 Received: from witch!lcanter.UUCP by frmug.fr.net; Thu, 17 Mar 94 05:29:18 +0100 Received: from witch.witchcraft.com (witch.witchcraft.com [198.30.130.2]) by cismsun.univ-lyon1.fr (8.6.7/8.6.6) with SMTP id FAA17873 for <nat@nataa.frmug.fr.net>; Thu, 17 Mar 1994 05:06:21 +0100 Received: by witch.witchcraft.com (5.65/1.35) id AA23994; Wed, 16 Mar 94 23:05:25 -0500 Received: by lcanter.win.net; Wed, 16 Mar 1994 21:09:11 Mailer:

WinNET Mail, v2.04 Message-Id: <2993@lcanter.win.net> Reply-To: cslaw@lcanter.win.net (Laurence A Canter) To: nat@nataa.frmug.fr.net Date: Wed, 16 Mar 1994 21:09:11 Subject: Re: U.S. Green Card Lottery - New Immigration Opportunity From: cslaw@lcanter.win.net (Laurence A Canter) Status: RO

Please note, the end of this message includes a biographical questionnaire. Please let us know if you do not receive the entire transmission.

Canter & Siegel 3333 E Camelback Road Ste 250 Phoenix AZ 85018 602 661 3911 (telephone) 602-451-7617 (fax) e-mail cslaw@indirect.com

WE CAN MAKE IT EASY TO APPLY AND INCREASE YOUR CHANCE OF WINNING ONE OF 55,000 GREEN CARDS AVAILABLE IN THE 1994 GREEN CARD LOTTERY.

What is the Green Card Lottery?

The Green Card Lottery is a program run by the United State Government to give away a certain number of Green Cards each year. In 1994, the number of Green Cards in the lottery is 55,000. The Green Card Lottery is completely legal in every way. What is unique about the lottery is that unlike other ways of getting Green Cards, you need no special qualification to apply. You need only have been born in one of the countries included in the program, If you win, in order to collect your Green Card, you must then show you have either a high school diploma, or at least two years of training or experience in a skilled job. You do not need to have a job offer.

THE TRUTH ABOUT THE GREEN CARD LOTTERY

FACT: The 1994 Green Card Lottery applies to people from almost all countries. In fact, only eight countries are excluded. They are: Mexico, India, China, Philippines, Dominican Republic, El Salvador, Jamaica and Korea. Four other countries are uncertain at this time. They are Canada, England, Taiwan, and Viet Nam. EVERYONE ELSE DEFINITELY QUALIFIES.

FACT: IF YOU WERE BORN IN ONE OF THE MANY QUALIFYING COUNTIES, YOU NEED NO OTHER QUALIFICATION TO WIN. YOU MAY BE PHYSICALLY INSIDE OR OUTSIDE THE U.S TO APPLY, IT DOES NOT MATTER. YOU MAY HAVE SOME OTHER TYPE OF U.S. VISA OR YOU MAY HAVE NO VISA AT ALL, IT DOES NOT MATTER. IT DOES NOT MATTER IF YOU ARE PRESENTLY TRYING TO GET A GREEN CARD IN SOME OTHER WAY. IT DOES NOT MATTER IF YOU ARE OR HAVE BEEN IN THE U.S. ILLEGALLY, AS LONG AS

YOU HAVE NOT BEEN DEPORTED WITHIN THE PAST FIVE YEARS, OR CONVICTED OF SERIOUS CRIMES. YOU CAN STILL ENTER THE LOTTERY.

FACT: The Green Card Lottery is run by the United States government and it is completely legal.

FACT: The Green Cards you can win in the Green Card lottery are exactly the same as all other Green Cards. They allow you to live and work legally in the United States. If you win and you are married, your spouse automatically gets a Green Card too. So do any unmarried children under age 21.

FACT: The countries eligible for the lottery change each year. Even though you are eligible in 1994, your country may not be on the list in 1995.

FACT: There is no official form. on which to apply for the lottery. Each application must be drawn up on a plain piece of paper according to strict specifications. The method of applying also must be carried out following preciseness.

FACT: Applying for the lottery has no effect of any kind on any other application you may make for U.S. residency, past, present or future. Applying for the lottery is completely safe.

FACT: There is no way of knowing exactly what your odds are of winning, because there is no way of telling how many people will apply this year, but last year, about 800,000 people filed qualifying applications for only 40,000 cards available. With these numbers your odds would be 1 out of 20.

FACT: You can put in only one application per person. If you put in more the computer will discover it and you will be disqualified.

FACT: This year's lottery is scheduled to take place in spring 1994. THERE IS A STRICT DEADLINE. IF YOUR APPLICATION DOES NOT ARRIVE DURING THE CORRECT TIME PERIOD, YOU CANNOT WIN.

FACT: LAST YEAR NEARLY 1/3 OF THE 1.1 MILLION APPLICATIONS SUBMITTED WERE THROWN OUT BECAUSE THEY WERE NOT DONE ACCORDING TO THE STRICT TECHNICAL REQUIREMENTS DEMANDED.

FACT: Everyone would like to know how to improve their chances to win the Green card lottery. THE TRUTH IS AN ATTORNEY CAN HELP INCREASE YOUR ODDS OF WINNING.

First, in certain cases only, you can as much as double or triple your chances by filing

several applications for the same family but using a different family member each time as the principal applicant. (This is not the same as submitting more than one application per person) A husband and wife may each be used as a principal applicant if each one qualifies individually. A child of a family may be given three chances by having one application filed with the child as principal and two others with each parent as principal. Please remember, however, that while a parent may include children on his or her application, children may not include parents. An attorney can see that your family's applications are properly handled.

Second, and most important, an attorney can guarantee that you don't make the same technical mistakes so many others have made. An attorney can insure that your application arrives at the right place, at the right time, and with no technical flaws that would cost you a real chance at winning the Green Card Lottery.

THE LAW FIRM OF CANTER & SIEGEL

The law firm of Canter and Siegel has been practicing Immigration law since 1981. In that time is has successfully acquired Green Cards and Visas for people from almost every country in the world. The firm offers a full range of Immigration services and handles all types of Immigration matters. The firm of Canter and Siegel practices only Immigration law and does not take cases in other areas. It has actively participated in every Green Card Lottery held since these programs began in 1987.

In 1989 Mr. Canter and Ms. Siegel co-authored a book for non-lawyers on the subject of Immigration called U.S. Immigration Made Easy. This book, now in its fourth edition, is fact becoming the major non-lawyer reference book in the field. Among the many excellent reviews received by this book, it has been highly recommended by the United States Information Agency and selected as one of the best books of the year by Library Journal. In 1992, Mr. Canter and Ms. Siegel co-authored another book on immigration, The Insider's Guide to Successful U.S. Immigration, published by Harper Collins.

Mr. Laurence A. Canter, a principal partner of the firm, received a law degree with honors from St. Mary's University where he served on law review. He then received an advanced law degree from Georgetown University, again near the top of his class. His undergraduate degree from the University of Arizona is in Spanish which he speaks fluently. He is a past national board of governors member of the American Immigration Lawyers Association. Ms. Martha S. Siegel received her undergraduate degree from Carnegie Mellon University and her law degree also from St. Mary's. In addition she has a Master of Library Science degree and is a graduate of the Library Internship Program

at Harvard University. Both lawyers are licensed to practice law in the state of Tennessee. Under federal regulations, they are authorized to practice immigration law in all fifty states. They do not accept other typess of cases.

LOTTERY SERVICES

The law firm of Canter & Siegel will take the information we receive from you and use it to form a technically perfect lottery application. If you are married and/or have children, they will be included on your application. Your application will be sent to the proper U.S. government agency, at the proper time You will receive confirmation from our firm that your application has been mailed. We will notify you immediately when we receive notice that you have won. When the lottery selections are over and all winners have been selected, we will notify those who have not won. We guarantee that your lottery application will be done perfectly and will get there at the right time. If you would like to have Canter & Siegel enter you in the 1994 Green Card Lottery, please fill out the enclosed Service Order and Questionnaire

FEES FOR ENTERING THE GREEN CARD LOTTERY (All payments must be in U.S. Funds)

The law firm of Canter and Siegel will represent you in the 1994 Green Card Lottery for Only $95 U.S.

This fee includes you, your spouse if you are married, and up to two children. (Remember, this is only for children under the age of 21 and unmarried. Children who are married or over age 21 must each send in separate applications). If you have more than two children, there is a fee of $15 U.S. for each additional child.

Double your chances for $50 U.S. more!

If you are married, you may double your chances by having us file separate applications for both you and your spouse. Our additional fee is only $50 U.S., for a total of $145 U.S.. Additional consultation with an attorney Many people would like to know what their alternatives are for obtaining a Green Card or Nonimmigrant Visa in the event they do not win the lottery. If you wish to have a private telephone consultation with an experienced immigration attorney on immigration matters other than the lottery, we offer this service for and additional $75 U.S. Simply indicate that you want a consultation on the enclosed Service Order Form and someone from the firm will contact you to arrange for an appointment.

1994 Green Card Lottery SERVICE ORDER & QUESTIONNAIRE

Please return this completed Questionnaire by Mail , Fax or e-mail to Canter & Siegel 3333 E Camelback Road, Ste 250 Phoenix AZ 85018 Fax: 602-451-7617 Phone: 602-661-3911 e-mail cslaw@indirect.com

( )YES, I would like Canter & Siegel to enter me in the Green Card Lottery

( )Please file one application for me and my family at a fee of 95 dollars U.S.

( )I have more than two children. Please charge me 15 dollars U.S. for each additional child.

( )I want to double my chances. Please file one application for me and another for my spouse for only 50 dollars U.S. more, a total of 145 dollars U.S.

( )I would like a telephone consultation with an attorney for a fee of 75 dollars U.S.

Total Amount

Enclosed is my check for U.S. Funds Charge my ( )Visa ( )Mastercard ( )American Express Card No. Exp. Date Signature

Name: (First, Middle, LAST)

Mailing Address:

If living inside the U.S., last city and country of residence:

Telephone Number: Fax Number:

Date of Birth (Month, Day, Year)

Place of Birth (City, State/Province, Country)

Did you graduate from High School? ( )

If no, what job skills do you have and how many years of training or experience for each skill

If you are married or have children under age 21 and you want them to receive Green Cards with you, for each complete the following:

Spouse:

Full Name:

Date & Place of Birth (Include City, State/Province, Country)

Children:

Full Name:

Date & Place of Birth (Include City, State/Province, Country)

Full Name:

Date & Place of Birth (Include City, State/Province, Country)

Service Agreement & Guarantee

Upon receiving your completed Service order form and payment, the law firm of Canter & Siegel agrees to prepare your registration, make sure it meets all technical requirements of the lottery program, and assure that it arrives at the U.S. Department of State during the official registration period. We will send you a confirmation after your registration has been filed, and will also notify you later on of the results. Winners will be able to get Green Cards beginning in October.

We guarantee to prepare your application in a technically perfect manner and to submit it on time so that it will be accepted and you will be given full consideration for receiving a Green Card. We cannot guarantee specific results. The odds for success will depend on how many people file proper applications, and what countries they come from.

If you are selected as a lottery winner, we will be happy to assist you in preparing your final green card application at an additional fee, or you may handle it on your own. This agreement covers the lottery registration procedure only.

To assist us in preparing your registration, please complete the questionnaire and return it to us.

Thank you for choosing Canter & Siegel to represent you in the 1994 Green Card Lottery.

_____

APÊNDICE B – Lista de Palavras - Estudo 1

A tabela 62 traz a listagem com as 100 palavras utilizada no Estudo 1. Ao todo estavam disponíveis 73.717 palavras.

Tabela 48: Lista Palavras Estudo 1

1º	size="4">How	26º	teaspoons	51º	size="5">Incest	76º	justive
2º	incidently	27º	tonight	52º	urkeyBlog	77º	ardner's
3º	leet	28º	naïvet	53º	odelle	78º	leaf
4º	leep	29º	artery	54º	mperrone@cs	79º	democrat
5º	leen	30º	eplica	55º	leese	80º	erhaegen
6º	financières	31º	bleeding	56º	leesa	81º	lead
7º	vocado	32º	junction	57º	taggering	82º	leere
8º	leek	33º	rginine	58º	sNine@sNine	83º	confusingly
9º	bashrc	34º	pamassassin-devel	59º	theonion	84º	hapless
10º	calea_law	35º	folder-hooks	60º	red">	85º	un-defangs
11º	pider's	36º	size=-1><ahref="http	61º	nears	86º	omestead
12º	leed	37º	artern	62º	domainnames	87º	oubakeur
13º	cfm">Ecobuilder	38º	higher-ranking	63º	leery	88º	aption
14º	squick	39º	ipage=qd&amp	64º	eyer's	89º	id=104548">PopupWar
15º	icking	40º	b>upto	65º	leas	90º	contests
16º	vincent@cunniffe	41º	leeth	66º	mailling	91º	smokes
17º	hassle	42º	size="2">Acts	67º	hands-off	92º	smoker
18º	aptist	43º	donkey	68º	holiday	93º	offense
19º	journeys	44º	roduktfamilien	69º	lear	94º	cicr_op
20º	yourmoney	45º	tandem	70º	value="Online	95º	f0S+xN3a64eZh
21º	notifications	46º	eb-based	71º	three-book	96º	zdnprmfnl
22º	irregular	47º	quality--and	72º	leap	97º	wheelchair
23º	adolescent	48º	n'est	73º	lean	98º	eachers
24º	color=#0000ff><b>Get	49º	size=2>Deepak	74º	tourinfo	99º	tandby
25º	smokin	50º	protections	75º	leak	100º	baseline

APÊNDICE C – Lista de Palavras - Estudo 2

A tabela 49 traz a listagem com as 100 palavras utilizada no Estudo 2. Ao todo estavam disponíveis 55.887 palavras.

Tabela 49: Lista Palavras Estudo 2

1º	frazier	26º	steps	51º	id's	76º	olarak
2º	d-link	27º	insecure	52º	stranger's	77º	mobius
3º	leed	28º	verbiage	53º	fourten	78º	mandrakeforum
4º	geoff	29º	leech	54º	atrocities	79º	before
5º	chemically	30º	vacuum	55º	namreh	80º	unsupervised
6º	leeds	31º	payed	56º	know"s	81º	alltogther
7º	sub-culture	32º	phisics	57º	mailto	82º	overflows
8º	abetting	33º	steph	58º	rewarding	83º	bunnymechanics
9º	freecolorprinters	34º	sayfam	59º	display	84º	reader's
10º	desvairada	35º	sprach	60º	knows	85º	verfuegung
11º	re-listing	36º	draws	61º	foomatic-compatible	86º	drauf
12º	!_in_reference	37º	stacked	62º	easier	87º	crying
13º	exert	38º	diligence	63º	!_in_console	88º	stems
14º	music-halls	39º	mkraid	64º	known	89º	lower-scoring
15º	auto-pack	40º	geode	65º	quotfinally	90º	directdsl
16º	toolbox	41º	drawn	66º	insure	91º	ls126a
17º	allowance	42º	liters	67º	bubblebutt	92º	northfield
18º	ineligible	43º	less-standard	68º	well-designed	93º	cxins
19º	un-needed	44º	savimbi	69º	leap	94º	beirne
20º	canada	45º	household	70º	one-to-one	95º	top-endgeräte
21º	payer	46º	newscast	71º	lean	96º	etruria
22º	monophonic	47º	!_in_highlight	72º	leak	97º	rightwingnuts
23º	expertise	48º	societal	73º	n'ai	98º	argue
24º	sunset	49º	brews	74º	leaf	99º	thievery
25º	mecca	50º	nearly-full	75º	lead	100º	developments

APÊNDICE D – Lista de Palavras - Estudo 3

A tabela 50 traz a listagem com as 100 palavras utilizada no Estudo 3 para o método DF, a tabela 51 traz as palavras utilizadas no método Chi-Quadrado e a tabela 52 traz as palavras consideradas pelo método MI. Ao todo estavam disponíveis 73.717 palavras.

Tabela 50: Lista Palavras Estudo 3 - Método DF

1º	the	26º	from	51º	list	76º	here
2º	and	27º	he	52º	email	77º	only
3º	com	28º	net	53º	about	78º	new
4º	width	29º	table	54º	cellspacing	79º	been
5º	www	30º	not	55º	cellpadding	80º	lists
6º	font	31º	will	56º	they	81º	sans-serif
7º	for	32º	cnet	57º	get	82º	div
8º	you	33º	can	58º	home	83º	what
9º	http	34º	his	59º	just	84º	information
10º	that	35º	our	60º	use	85º	face="Verdana
11º	td>	36º	all	61º	zdnet	86º	org
12º	height	37º	online	62º	which	87º	other
13º	nbsp	38º	was	63º	img	88º	some
14º	gif	39º	color	64º	listinfo	89º	there
15º	src="http	40º	lick	65º	rial	90º	now
16º	with	41º	but	66º	any	91º	mail
17º	tr>	42º	br>	67º	a><	92º	than
18º	your	43º	clickthru	68º	like	93º	mailman
19º	this	44º	html	69º	would	94º	them
20º	href="http	45º	has	70º	time	95º	who
21º	td	46º	more	71º	their	96º	mailing
22º	size	47º	b><	72º	ou	97º	when
23º	are	48º	out	73º	face="Arial	98º	how
24º	have	49º	elvetica	74º	a><br	99º	make
25º	border	50º	one	75º	people	100º	lockergnome

Tabela 51: Lista Palavras Estudo 3 - Método Chi-Quadrado

1º	our	26º	listinfo	51º	offer	76º	cellspacing
2º	href="http	27º	tr>	52º	meta	77º	content="Microsoft
3º	font	28º	ere	53º	align="center"><font	78º	cellPadding
4º	your	29º	height	54º	money	79º	mortgage
5º	size	30º	table	55º	offers	80º	reply
6º	removed	31º	email	56º	credit	81º	cellSpacing
7º	width	32º	wish	57º	ou	82º	p><font
8º	border	33º	href="mailto	58º	head	83º	hours
9º	remove	34º	align="center	59º	within	84º	rontPage
10º	color	35º	b><font	60º	redit	85º	subject
11º	a><	36º	ur	61º	removal	86º	index
12º	b><	37º	below	62º	future	87º	will
13º	br>	38º	you	63º	hank	88º	address
14º	please	39º	mailings	64º	sans-serif	89º	lick
15º	wrote	40º	but	65º	contact	90º	internet
16º	body	41º	name	66º	src="http	91º	opportunity
17º	receive	42º	rial	67º	form	92º	illion
18º	content="text	43º	td	68º	p>	93º	ocument
19º	font><font	44º	http	69º	title	94º	center
20º	html	45º	bgcolor	70º	marketing	95º	inux
21º	face="Arial	46º	business	71º	link	96º	send
22º	lease	47º	i><	72º	rder	97º	m
23º	td>	48º	div	73º	cellpadding	98º	ortgage
24º	mailman	49º	elvetica	74º	charset=iso	99º	content="FrontPage
25º	nbsp	50º	face="Verdana	75º	ree	100º	name="ProgId

Tabela 52: Lista Palavras Estudo 3 - Método MI

1º	font	26º	the	51º	cnet	76º	said
2º	nbsp	27º	wrote	52º	sans-serif"><font	77º	emple
3º	color	28º	business	53º	pt	78º	spam
4º	you	29º	span	54º	listinfo	79º	file
5º	size	30º	money	55º	remove	80º	lease
6º	com	31º	www	56º	p><font	81º	will
7º	your	32º	a><br	57º	meta	82º	send
8º	face="Arial	33º	face=Arial	58º	i><	83º	tr><td
9º	br>	34º	org	59º	li>	84º	m
10º	elvetica	35º	tr>	60º	removed	85º	below
11º	font><font	36º	ffffff	61º	border="0"><br	86º	align="left
12º	div	37º	b><font	62º	blockquote><font	87º	html
13º	face="Verdana	38º	email	63º	helvetica	88º	orders
14º	sans-serif	39º	receive	64º	ere	89º	face="Times
15º	rial	40º	align=center	65º	eneva	90º	rants
16º	our	41º	that	66º	cm	91º	redit
17º	http	42º	but	67º	graphics	92º	ug
18º	gif	43º	inux	68º	credit	93º	e-mails
19º	blockquote	44º	name	69º	style	94º	sans
20º	online	45º	face="Verdana"><font	70º	href="mailto	95º	some
21º	b><	46º	align="center"><font	71º	clickthru	96º	ew
22º	src="http	47º	mailman	72º	please	97º	rder
23º	option	48º	body	73º	content="text	98º	ed
24º	lick	49º	face="Tahoma	74º	value	99º	mailings
25º	align="center	50º	center	75º	img	100º	mail

APÊNDICE E – Lista de Palavras - Estudo 4

A tabela 53 traz a listagem com as 100 palavras utilizada no Estudo 4 para o método DF, a tabela 54 traz as palavras utilizadas no método Chi-Quadrado e a tabela 55 traz as palavras consideradas pelo método MI. Ao todo estavam disponíveis 55.887 palavras.

Tabela 53: Lista Palavras Estudo 4 - Método DF

1º	the	26º	!_in_helvetica	51º	!_in	76º	click
2º	and	27º	will	52º	only	77º	than
3º	!_in_font	28º	!_in_border	53º	any	78º	who
4º	you	29º	all	54º	!_in_name	79º	!_in_type
5º	!_in_size	30º	can	55º	which	80º	also
6º	for	31º	but	56º	there	81º	them
7º	!_in_face	32º	was	57º	!_in_arial	82º	business
8º	that	33º	our	58º	time	83º	mailing
9º	!_LINK	34º	more	59º	people	84º	then
10º	this	35º	list	60º	like	85º	message
11º	!_in_href	36º	one	61º	would	86º	money
12º	!_in_img	37º	email	62º	use	87º	over
13º	!_in_src	38º	get	63º	information	88º	linux
14º	!_in_height	39º	!_in_sans-serif	64º	their	89º	web
15º	!_in_width	40º	has	65º	how	90º	work
16º	your	41º	out	66º	when	91º	its
17º	!_in_color	42º	!_in_alt	67º	!_PORCENTAGEM	92º	into
18º	with	43º	they	68º	!_in_input	93º	address
19º	are	44º	just	69º	some	94º	these
20º	from	45º	free	70º	been	95º	want
21º	have	46º	about	71º	other	96º	wrote
22º	not	47º	new	72º	don't	97º	may
23º	!_BIGTEXT	48º	now	73º	please	98º	internet
24º	!_EMAIL	49º	what	74º	it's	99º	see
25º	!_MONEY	50º	here	75º	make	100º	most

Tabela 54: Lista Palavras Estudo 4 - Método Chi-Quadrado

1º	click	26º	!_LINK	51º	i'm	76º	emails
2º	!_in_href	27º	credit	52º	future	77º	opt-in
3º	!_in_font	28º	but	53º	purchase	78º	best
4º	our	29º	form	54º	cash	79º	state
5º	!_in_size	30º	!_in_sans-serif	55º	dear	80º	receiving
6º	!_in_color	31º	money	56º	thank	81º	obligation
7º	!_in_face	32º	fill	57º	date	82º	url
8º	please	33º	business	58º	call	83º	low
9º	your	34º	offer	59º	company	84º	lowest
10º	removed	35º	!_in_width	60º	within	85º	cost
11º	receive	36º	!_in_height	61º	rates	86º	financial
12º	here	37º	you	62º	e-mail	87º	guarantee
13º	wrote	38º	guaranteed	63º	insurance	88º	interest
14º	remove	39º	marketing	64º	today	89º	life
15º	wish	40º	!_in_border	65º	opportunity	90º	!_in_subject
16º	free	41º	name	66º	income	91º	online
17º	!_MONEY	42º	offers	67º	!_in_arial	92º	per
18º	below	43º	mortgage	68º	!_EMAIL	93º	investment
19º	!_PORCENTAGEM	44º	removal	69º	linux	94º	i've
20º	!_in_helvetica	45º	address	70º	special	95º	million
21º	!_in_img	46º	will	71º	!_in_roman	96º	send
22º	!_in_src	47º	contact	72º	hundreds	97º	prices
23º	email	48º	hours	73º	group	98º	aug
24º	mailings	49º	visit	74º	yourself	99º	amp
25º	reply	50º	dollars	75º	order	100º	professional

Tabela 55: Lista Palavras Estudo 4 - Método MI

1º	!_in_font	26º	that	51º	marketing	76º	program
2º	!_in_color	27º	credit	52º	send	77º	income
3º	!_in_size	28º	linux	53º	insurance	78º	was
4º	!_in_face	29º	!_EMAIL	54º	remove	79º	addresses
5º	you	30º	!_in_height	55º	e-mails	80º	users
6º	your	31º	!_in_width	56º	perl	81º	!_in_ptsize
7º	!_in_sans-serif	32º	!_in_input	57º	!_in_type	82º	reply
8º	!_MONEY	33º	grants	58º	file	83º	guide
9º	!_LINK	34º	!_in_name	59º	aug	84º	i've
10º	click	35º	report	60º	!_in_black	85º	server
11º	the	36º	name	61º	mailings	86º	files
12º	money	37º	removed	62º	some	87º	this
13º	free	38º	!_in_img	63º	dollars	88º	temple
14º	!_in_helvetica	39º	!_in_src	64º	cash	89º	xml
15º	!_in_blockquote	40º	here	65º	date	90º	home
16º	our	41º	i'm	66º	address	91º	kingdom
17º	wrote	42º	will	67º	wish	92º	amp
18º	business	43º	!_in_roman	68º	guaranteed	93º	toner
19º	receive	44º	orders	69º	url	94º	they
20º	!_in_arial	45º	below	70º	fax	95º	form
21º	email	46º	!_in	71º	group	96º	save
22º	!_BIGTEXT	47º	offer	72º	spam	97º	rates
23º	please	48º	mortgage	73º	rpm	98º	cnet
24º	but	49º	said	74º	!_PORCENTAGEM	99º	purchase
25º	order	50º	!_in_new	75º	e-mail	100º	million

APÊNDICE F – Lista de Palavras - Estudo 5

A tabela 56 traz a listagem com as 100 palavras utilizada no Estudo 5 para o método DF, a tabela 57 traz as palavras utilizadas no método Chi-Quadrado e a tabela 58 traz as palavras consideradas pelo método MI. Ao todo estavam disponíveis 51.365 palavras.

Tabela 56: Lista Palavras Estudo 5 - Método DF

1º	language	26º	word	51º	available	76º	day
2º	university	27º	follow	52º	process	77º	provide
3º	!_MONEY	28º	papers	53º	page	78º	department
4º	linguistic	29º	edu	54º	each	79º	international
5º	information	30º	call	55º	speech	80º	most
6º	address	31º	system	56º	case	81º	speaker
7º	one	32º	report	57º	text	82º	money
8º	conference	33º	e-mail	58º	web	83º	those
9º	send	34º	abstract	59º	th	84º	linguist
10º	order	35º	www	60º	!_NUMERO_SUBJECT	85º	structure
11º	please	36º	interest	61º	many	86º	over
12º	english	37º	book	62º	student	87º	computer
13º	include	38º	study	63º	!_PORCENTAGEM	88º	discussion
14º	work	39º	theory	64º	question	89º	syntax
15º	workshop	40º	submission	65º	state	90º	must
16º	mail	41º	form	66º	issue	91º	area
17º	paper	42º	first	67º	science	92º	discourse
18º	email	43º	receive	68º	author	93º	present
19º	http	44º	number	69º	between	94º	here
20º	program	45º	session	70º	contact	95º	site
21º	name	46º	linguistics	71º	registration	96º	example
22º	fax	47º	two	72º	analysis	97º	need
23º	our	48º	com	73º	write	98º	different
24º	list	49º	free	74º	usa	99º	reference
25º	research	50º	grammar	75º	copy	100º	john

Tabela 57: Lista Palavras Estudo 5 - Método Chi-Quadrado

1º	remove	26º	easy	51º	watch	76º	off
2º	free	27º	yourself	52º	offer	77º	friend
3º	language	28º	day	53º	wait	78º	net
4º	money	29º	every	54º	ever	79º	week
5º	click	30º	bulk	55º	amaze	80º	pay
6º	university	31º	mailing	56º	toll	81º	xxx
7º	sell	32º	dollar	57º	month	82º	profitable
8º	today	33º	com	58º	bonus	83º	linguistics
9º	linguistic	34º	want	59º	live	84º	security
10º	market	35º	best	60º	zip	85º	package
11º	business	36º	cash	61º	huge	86º	marketing
12º	product	37º	!_MONEY	62º	dream	87º	simply
13º	million	38º	earn	63º	english	88º	enter
14º	save	39º	hundred	64º	investment	89º	fantastic
15º	company	40º	check	65º	success	90º	ship
16º	!_PORCENTAGEM	41º	cost	66º	here	91º	excite
17º	income	42º	profit	67º	anywhere	92º	aol
18º	advertise	43º	customer	68º	receive	93º	keep
19º	our	44º	over	69º	credit	94º	back
20º	internet	45º	service	70º	financial	95º	spend
21º	thousand	46º	hour	71º	start	96º	everything
22º	win	47º	fun	72º	secret	97º	hit
23º	guarantee	48º	sale	73º	mlm	98º	never
24º	buy	49º	online	74º	advertisement	99º	papers
25º	purchase	50º	yours	75º	fresh	100º	step

Tabela 58: Lista Palavras Estudo 5 - Método MI

1º	!_MONEY	26º	day	51º	cost	76º	investment
2º	language	27º	com	52º	financial	77º	ship
3º	report	28º	sell	53º	session	78º	bonus
4º	money	29º	advertise	54º	buy	79º	purchase
5º	order	30º	income	55º	earn	80º	linguist
6º	university	31º	send	56º	software	81º	speaker
7º	free	32º	credit	57º	today	82º	advertisement
8º	business	33º	click	58º	site	83º	pay
9º	our	34º	every	59º	paper	84º	analysis
10º	mail	35º	start	60º	multi-level	85º	structure
11º	internet	36º	name	61º	live	86º	web
12º	email	37º	papers	62º	guarantee	87º	author
13º	receive	38º	want	63º	directory	88º	legal
14º	remove	39º	win	64º	research	89º	aol
15º	market	40º	week	65º	sale	90º	opportunity
16º	program	41º	service	66º	submission	91º	mailing
17º	bulk	42º	cash	67º	link	92º	need
18º	product	43º	edu	68º	offer	93º	capitalfm
19º	conference	44º	save	69º	success	94º	month
20º	list	45º	company	70º	thousand	95º	net
21º	address	46º	letter	71º	dollar	96º	step
22º	million	47º	easy	72º	science	97º	down
23º	check	48º	best	73º	hour	98º	profit
24º	over	49º	linguistics	74º	study	99º	each
25º	english	50º	floodgate	75º	speech	100º	card

APÊNDICE G – Lista de Palavras - Estudo 6

A tabela 59 traz a listagem com as 100 palavras utilizada no Estudo 6 para o método DF, a tabela 60 traz as palavras utilizadas no método Chi-Quadrado e a tabela 61 traz as palavras consideradas pelo método MI. Ao todo estavam disponíveis 246.797 palavras.

Tabela 59: Lista Palavras Estudo 6 - Método DF

1º	the	26º	not	51º	per	76º	his
2º	!_in_font	27º	from	52º	here	77º	code
3º	and	28º	have	53º	list	78º	!_PORCENTAGEM
4º	!_in_size	29º	!_in_alt	54º	!_in_font-weight	79º	vous
5º	!_in_href	30º	!_in_sans-serif	55º	about	80º	how
6º	!_MONEY	31º	!_in_helvetica	56º	!_in_bold	81º	time
7º	!_in_face	32º	our	57º	would	82º	news
8º	you	33º	!_in_target	58º	pills	83º	see
9º	!_in_color	34º	all	59º	!_NUMERO_SUBJECT	84º	said
10º	!_BIGTEXT	35º	will	60º	only	85º	just
11º	for	36º	!_in_font-size	61º	now	86º	read
12º	!_in_img	37º	but	62º	like	87º	!_in_hspace
13º	!_in_src	38º	was	63º	out	88º	mailing
14º	your	39º	desjardins	64º	there	89º	their
15º	that	40º	more	65º	which	90º	save
16º	this	41º	can	66º	any	91º	!_in_sans
17º	with	42º	!_in	67º	votre	92º	than
18º	!_in_height	43º	!_in_text-decoration	68º	get	93º	!_in_11px
19º	!_in_width	44º	!_in_none	69º	they	94º	then
20º	!_LINK	45º	may	70º	!_HIDEWORDS	95º	!_in_vspace
21º	!_in_border	46º	new	71º	use	96º	other
22º	!_EMAIL	47º	one	72º	price	97º	need
23º	are	48º	has	73º	!_in_class	98º	been
24º	!_in_arial	49º	please	74º	some	99º	px
25º	!_in_style	50º	what	75º	when	100º	should

Tabela 60: Lista Palavras Estudo 6 - Método Chi-Quadrado

1º	!_EMAIL	26º	using	51º	alternative	76º	additional
2º	list	27º	!_in_color	52º	function	77º	set
3º	mailing	28º	samba	53º	from	78º	revision
4º	wrote	29º	perl	54º	the	79º	error
5º	code	30º	use	55º	new	80º	privacy
6º	!_in_font	31º	!_HIDEWORDS	56º	package	81º	email
7º	posting	32º	log	57º	example	82º	network
8º	minimal	33º	modified	58º	money	83º	should
9º	reproducible	34º	data	59º	does	84º	following
10º	self-contained	35º	this	60º	test	85º	viagra
11º	commented	36º	com	61º	rights	86º	files
12º	read	37º	svn	62º	rev	87º	sunday
13º	guide	38º	i'm	63º	help	88º	saturday
14º	!_in_face	39º	file	64º	changeset	89º	return
15º	!_LINK	40º	commands	65º	deleted	90º	may
16º	!_in_href	41º	change	66º	reserved	91º	speakup
17º	unsubscribe	42º	subject	67º	subscribed	92º	some
18º	!_in_size	43º	which	68º	!_NUMERO_SUBJECT	93º	provided
19º	provide	44º	commit	69º	sent	94º	!_in_sans-serif
20º	!_BIGTEXT	45º	html	70º	think	95º	hash
21º	please	46º	author	71º	alerts	96º	seems
22º	version	47º	date	72º	your	97º	trying
23º	e-mail	48º	linux	73º	our	98º	view
24º	but	49º	how	74º	utc	99º	variable
25º	message	50º	websvn	75º	jun	100º	null

Tabela 61: Lista Palavras Estudo 6 - Método MI

1º	!_in_font	26º	!_in_sans	51º	les	76º	!_in_link
2º	!_in_size	27º	px	52º	!_in_width	77º	!_LINK
3º	!_in_face	28º	mailing	53º	money	78º	retail
4º	!_in_style	29º	!_in_sans-serif	54º	adobe	79º	avis
5º	!_in_font-size	30º	!_in_vspace	55º	perl	80º	degc
6º	desjardins	31º	wrote	56º	!_HIDEWORDS	81º	e-mail
7º	your	32º	viagra	57º	guide	82º	pour
8º	!_in_none	33º	!_in	58º	!_in_arial	83º	!_in_ptsize
9º	!_in_text-decoration	34º	mg	59º	return	84º	!_in_lang
10º	!_BIGTEXT	35º	minimal	60º	que	85º	groupe
11º	you	36º	save	61º	struct	86º	file
12º	!_in_bold	37º	posting	62º	modified	87º	arial
13º	pills	38º	!_in_font-family	63º	int	88º	!_in_family
14º	!_in_font-weight	39º	self-contained	64º	cialis	89º	rev
15º	!_EMAIL	40º	price	65º	read	90º	tout
16º	votre	41º	!_in_src	66º	transactions	91º	institution
17º	our	42º	transaction	67º	utc	92º	!_in_padding
18º	!_MONEY	43º	commented	68º	!_in_10px	93º	cnn
19º	vous	44º	data	69º	sid	94º	quality
20º	!_in_color	45º	!_in_lid	70º	nous	95º	if
21º	per	46º	!_in_height	71º	!_in_img	96º	men
22º	!_in_11px	47º	pas	72º	the	97º	une
23º	code	48º	char	73º	const	98º	version
24º	!_in_12px	49º	item	74º	null	99º	!_in_alt
25º	list	50º	!_in_helvetica	75º	samba	100º	online

APÊNDICE H – Lista de Palavras - Estudo 7

A tabela 62 traz a listagem com as 100 palavras utilizada no Estudo 7.

Tabela 62: Lista Palavras Estudo 7

1º	the	26º	redhat	51º	can	76º	irish
2º	and	27º	list	52º	kickstart	77º	information
3º	for	28º	there	53º	more	78º	implemented
4º	!_HIDEWORDS	29º	well	54º	what	79º	un)subscription
5º	!_LINK	30º	would	55º	file	80º	floppy
6º	you	31º	know	56º	which	81º	maintainer
7º	that	32º	but	57º	group	82º	boot
8º	this	33º	linux	58º	server	83º	billions
9º	!_EMAIL	34º	not	59º	sequences	84º	doesn't
10º	have	35º	how	60º	open	85º	canada
11º	with	36º	our	61º	metacity	86º	while
12º	are	37º	new	62º	settings	87º	some
13º	from	38º	take	63º	libertarianism	88º	window
14º	all	39º	was	64º	dollars	89º	source
15º	just	40º	mpeg	65º	japanese	90º	world
16º	use	41º	has	66º	going	91º	packages
17º	it's	42º	had	67º	america	92º	telecom
18º	get	43º	using	68º	where	93º	any
19º	i'm	44º	one	69º	unseen	94º	about
20º	streaming	45º	north	70º	video	95º	did
21º	card	46º	i've	71º	own	96º	sawfish
22º	machine	47º	off	72º	other	97º	time
23º	many	48º	like	73º	could	98º	try
24º	out	49º	don't	74º	does	99º	users
25º	network	50º	when	75º	people	100º	say