UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS GRADUAÇÃO EM
CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO

Estudo e Avaliação de Classificadores para um Sistema Anti-Spam

Marcelo Vinícius Cysneiros Aragão

Itajubá, junho de 2018

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

Marcelo Vinícius Cysneiros Aragão

Estudo e Avaliação de Classificadores para um Sistema Anti-Spam

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: Prof. Dr. Otávio Augusto Salgado Carpinteiro
Coorientador: Prof. Dr. Edmilson Marmo Moreira

Junho de 2018
Itajubá - MG

Agradecimentos

Agradeço aos amigos e familiares pela paciência, tolerância e apoio emocional durante esta jornada turbulenta e gratificante.

Também registro minha gratidão aos colegas e mentores que me ouviram, criticaram, opinaram e, acima de tudo, me ajudaram a manter a cabeça erguida e a mente aberta, de modo que pudesse fazer o melhor trabalho possível, diante das circunstâncias às quais fui sujeitado nos últimos anos.

Finalmente, um saudoso agradecimento com sabor de dedicatória ao meu avô Giuseppe (*in memoriam*), por ter me mostrado que ser "mestre" não consiste apenas em receber um título por cumprir determinados requisitos acadêmicos, mas também em manifestar diariamente a incessável vontade de aprender e consequente responsabilidade de ensinar.

Grande astro! Que seria da tua felicidade se te faltassem aqueles a quem iluminas?

Friedrich Wilhelm Nietzsche

Resumo

O advento das redes de dados móveis vem popularizando o acesso à Internet, aumentando o número de usuários e o volume de troca de *e-mails*. Desde 2005, porém, pelo menos metade deste tráfego é composto por *spam*, ou seja, mensagens não-solicitadas enviadas em massa a fim de promover um produto ou serviço, podendo incluir anúncios, esquemas de pirâmide, doações, correntes, entre outros.

Este tráfego massivo de mensagens indesejadas traz prejuízos aos usuários, tais como o uso excessivo e desnecessário da largura de banda, impacto ambiental, perda de produtividade, exposição de conteúdo impróprio a públicos sensíveis, prejuízos financeiros e legais e redução da efetividade de propagandas legítimas.

Uma técnica largamente empregada por servidores de *e-mail*, a fim de lidar com este problema são as *listas negras*, um banco de dados em tempo real que determina se um endereço IP está enviando *e-mails* que podem ser considerados *spam*. Este tipo de sistema apresenta desvantagens, tais como código proprietário, parcialidade questionável, limitações dos protocolos IPv4/IPv6 e lentidão na classificação.

Outra abordagem ao problema de classificação de *e-mails* consiste em usar os dados contidos nas próprias mensagens e um modelo de aprendizado de máquina, a fim de construir um sistema de decisão automático, inteligente, imparcial e adaptativo, capaz de identificar padrões e se adaptar às características comuns a ambos os tipos de *e-mail*, apresentando alta precisão de classificação.

Portanto, o objetivo deste trabalho é o estudo e a aplicação de métodos de aprendizado de máquina à classificação de *e-mails*, a fim de obter modelos utilizáveis em sistemas anti-*spam* do mundo real. Para isto, são também abordadas técnicas de seleção de características e redução de dimensionalidade, necessária e complementar, respectivamente, ao pré-processamento dos dados utilizados pelos algoritmos.

Após realizar experimentos em diversos *corpora*, os resultados confirmaram que modelos de aprendizado de máquina, aliados a métodos de seleção de características baseados em informação mútua e às técnicas de redução de dimensionalidade por correlação, podem ser aplicados com sucesso à categorização de *e-mails*, detectando *spams* com alto grau de confiabilidade para o usuário.

Palavras-chave: aprendizado de máquina. categorização de texto. anti-spam.

Abstract

The advent of mobile data networks has been popularizing Internet access, increasing the number of *e-mail* users and the volume of exchange of these messages. Since 2005, however, at least half of this traffic is made up of *spam*. i.e., unsolicited messages sent in bulk to promote a product or service, which may include advertisements, pyramid schemes, donations, chains, among others.

This massive traffic of unwanted messages causes harm to users, such as excessive and unnecessary bandwidth use, environmental impact, loss of productivity, exposure of sensitive audiences to inappropriate content, financial and legal damage, and reduction of the effectiveness of legitimate advertisements.

One technique widely employed by *e-mail* servers to deal with this problem is *blacklisting*, a real-time database that determines whether an IP address is sending *e-mail* that may be considered *spam*. This type of system has disadvantages such as prorietary code, questionable partiality, limitations of IPv4/IPv6 protocols and slowness in classification.

Another approach to the problem of *e-mail* classification is to use the data contained in the messages themselves and a machine learning method to build an automatic, intelligent, unbiased and adaptive decision system capable of identifying patterns and adapting features common to both types of *e-mail*, presenting high classification accuracy.

Therefore, the objective of this work is the study and application of machine learning methods to the classification of *e-mails*, in order to obtain models usable in real-world anti-*spam* systems. For this, techniques of characteristicsselection of and dimensionality reduction, necessary and complementary, respectively, to the pre-processing of the data used by the algorithms will also be approached.

After performing experiments on several *corpora*, the results confirmed that machine learning models, coupled with mutual information-based characteristics selection methods and correlation-based dimensionality reduction techniques, can be applied successfully to the categorization of *e-mails*, detecting *spam* with a high degree of reliability for the user.

Keywords: machine learning. text categorization. anti-spam.

Sumário

Lista de Figuras

Lista de Tabelas

Glossário

1D-LBP	*One-Dimension Linear Binary Pattern*
AB-M1	*Boosted Reduced-Error Pruning Tree*
ADTrees	*Alternating Decision Trees*
AODE	*Averaged One-Dependence Estimators*
API	*Application Programming Interface*
ASCII	*American Standard Code for Information Interchange*
AT&T	*American Telephone and Telegraph*
BN	*BayesNet*
BNS	*Bi-Normal Separation*
CFS	*Correlation-based Feature Selection*
CUDA	*Compute Unified Device Architecture*
DF	*Document Frequency*
DNS	*Domain Name System*
DNSBL	*Domain Name System-based Blackhole List*
EAF	*Empirical Attainment Function*
EANT2	*Evolutionary Acquisition of Neural Topologies, Version 2*
ENORA	*Evolutionary NOn-dominated Radial slots based Algorithm*
EQM	Erro Quadrático Médio
EUA	Estados Unidos da América
FD	Frequência no Documento
FLDA	*Fisher Linear Discriminant Analysis*
FRES	*Flesch Reading Ease Score*
FT	*Functional Tree*
GNARL	*GeNeralized Acquisition of Recurrent Links*

GPESC	Grupo de Pesquisas em Engenharia de Sistemas e de Computação
GPU	*Graphics Processing Unit*
GRBF	*Generalized Radial Basis Functions*
HTML	*HyperText Markup Language*
IB	*Information Bottleneck*
ID3	*Iterative Dichotomiser 3*
IG	*Information Gain*
IP	*Internet Protocol*
IPv4	*Internet Protocol version 4*
IPv6	*Internet Protocol version 6*
ISP	*Internet Service Provider*
KNN	*k-Nearest Neighbors*
L-SVM	*Linear Support Vector Machine*
LMS	*Least Mean Squares*
LNO	*Largest Number of Objects*
LOF	*Local Outlier Factor*
LTS	*Largest Total Strength*
LVF	*Las Vegas Filter*
MAP	*Maximum A Posteriori*
MCC	*Matthews Correlation Coefficient*
MCP	*McCulloch–Pitts*
MFD	*Most Frequent Decision*
MI	*Mutual Information*
MIME	*Multipurpose Internet Mail Extensions*
MIT	*Massachusetts Institute of Technology*
MLP	*Multilayer Perceptron*
NB	*Naïve Bayes*
NEAT	*NeuroEvolution of Augmenting Topologies*
NL-SVM	*Non-Linear Support Vector Machine*
NSA-PSO	*Negative Selection Algorithm - Particle Swarm Optimization*

NSGA-II	*Nondominated Sorting Genetic Algorithm II*
Open-MaLBAS	*Open Machine Learning-Based Anti-Spam*
P-SVM	*Polynomial Support Vector Machine*
PCA	*Principal Component Analysis*
RBF	*Radial Basis Function*
RDF	*Resource Description Framework*
REPT	*Reduced-Error Pruning Tree*
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
RT	*Random Tree*
SBL	*Spamhaus Block List*
SLP	*Single Layer Perceptron*
SMOG	*Simple Measure of Gobbledygook*
SOM	*Self-Organizing Map*
SPEA2	*Strength Pareto Evolutionary Algorithm 2*
SSF	*Simplified Silhouette Filter*
STA	Servidor Técnico-Administrativo
SVC	*Support Vector Clustering*
SVM	*Support Vector Machine*
SVM-RFE	*Support Vector Machine Recursive Feature Elimination*
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
TREC	*Text REtrieval Conference*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
UCE	*Unsolicited Commercial E-mail*
UNIFEI	Universidade Federal de Itajubá
UTF-8	*8-bit Unicode Transformation Format*
WEKA	*Waikato Environment for Knowledge Analysis*

1 Introdução

1.1 Contexto da pesquisa e justificativas

Nos dias de hoje, com o advento das redes de dados móveis, um número cada vez maior de pessoas possui acesso à Internet de qualquer lugar. Seja no trabalho, em casa ou até mesmo no trânsito, é possível receber e enviar *e-mails* a qualquer momento. Isto implica em um aumento na quantidade de usuários de *e-mail* e, consequentemente, na quantidade de trocas de mensagens, conforme pode ser confirmado nos gráficos da figuras 1 e 2 (RADICATI, 2018), respectivamente.

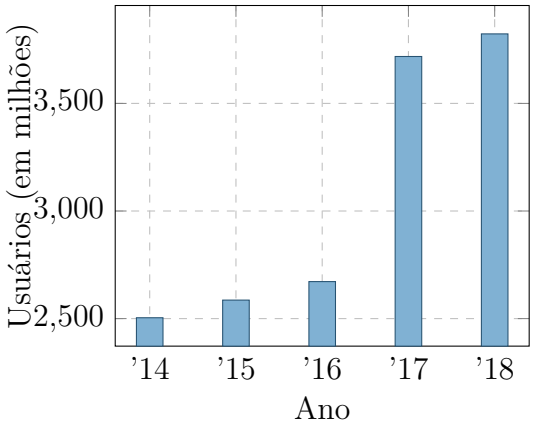


Figura 1: Usuários de *E-mail*.

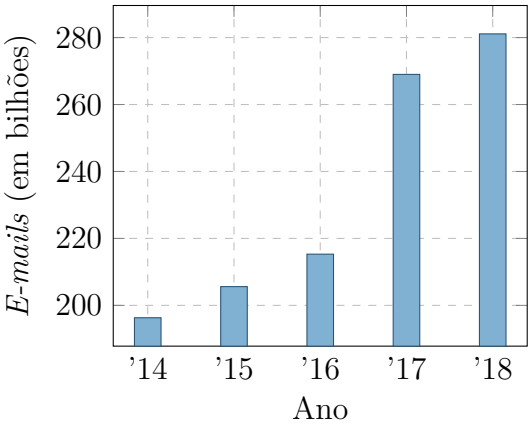


Figura 2: Tráfego Diário de *E-mails*.

Infelizmente não são somente a quantidade de usuários e a facilidade de acesso que vêm aumentando. De acordo com Symantec Corporation (2005), Symantec Corporation (2012) e Symantec Corporation (2018), pelo menos metade do tráfego de *e-mails* desde 2005 é composta por *spam* (vide figura 3), ou seja, mensagens

não solicitadas cujo destinatário não é um conhecido, cliente ou um membro voluntariamente inscrito em uma lista de distribuição. Neste mesmo sentido, é comum fazer referência a *e-mails* legítimos (não *spam*) por meio do termo *ham*.

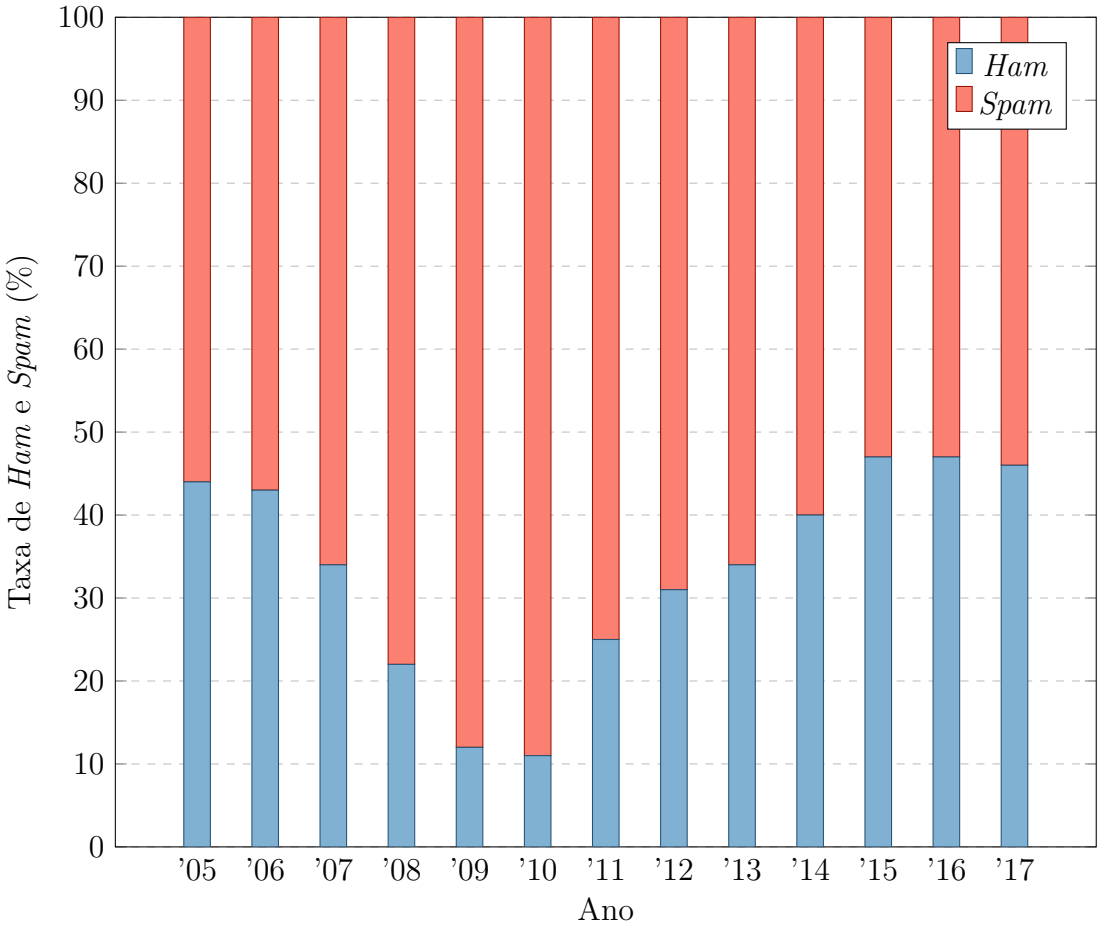


Figura 3: Percentual de *Ham* e *Spam* ao Ano.

Geralmente, *spams* são enviados a um grande número de pessoas para promover um produto ou serviço e, por esta razão, muitas vezes são chamados de *e-mail* comercial não solicitados (UCE, do inglês *Unsolicited Commercial E-mail*). Isso inclui anúncios, esquemas de pirâmide, doações, correntes de mensagens, *e-mail* político, assessoria no mercado de ações e avisos únicos.

Este tráfego massivo de mensagens indesejadas provoca diversas consequências negativas, podendo ser citadas:

- Uso excessivo da largura de banda, provocando lentidão de ISPs (do inglês *Internet Service Provider*) e sobrecarga de servidores de *e-mail*.

- Impacto ambiental: considerando o custo energético para o processamento de *spam*, a energia gasta poderia ser utilizada para outros fins (McAfee Inc.; ICF International, 2009).

- Perda de produtividade: o recebimento de *spam* diminui a produtividade de todos que utilizam *e-mails* como ferramenta de trabalho, pois um tempo extra é dedicado à tarefa de recebimento e leitura, podendo também uma mensagem importante ser apagada por engano, ignorada ou lida com atraso dado o volume de *spams* recebidos. Esta perda também é diretamente associada a prejuízos financeiros.

- 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, podendo até expor crianças a material inadequado.

- Prejuízos financeiros e legais 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.

- Redução da efetividade de propagandas legítimas.

Diante dos problemas apresentados, nota-se a importância de filtrar este tipo de mensagem. Uma solução resultaria em um alívio nos sistemas de telecomunicações e em melhor experiência do usuário na utilização de seu *e-mail*.

Um usuário experiente de computador consegue facilmente identificar um *spam*, seja por meio de seu conteúdo (palavras-chave, *links* e/ou imagens), seja por

mera vigilância epistêmica[1](SPERBER et al., 2010) adquirida com o tempo e, possivelmente, com experiências ruins com este tipo de mensagem (exemplo: ser alvo de algum tipo de fraude). Entretanto, para um computador, este processo não é tão trivial. Isto porque o cérebro humano, devido à sua natureza, trabalha de forma altamente paralelizada e possui grande capacidade de aprendizado. Já o computador apresenta comportamento linear, e a identificação de *spams* envolve a criação de regras complexas e algoritmos custosos para processamento, análise e categorização de texto.

Este processo de categorização é normalmente feito utilizando modelos de aprendizado de máquina, sejam estes supervisionados (com amostras rotuladas) ou não-supervisionados (com amostras de treinamento sem classificação). Estes modelos são capazes de identificar padrões e de se ajustar a características comuns a ambos os tipos de *e-mail*, resultando em classificações cada vez mais precisas, com o decorrer das épocas de treinamento. Entretanto, cada modelo apresenta defeitos e qualidades particulares que influenciam na aplicabilidade em sistemas de classificação e processos decisórios em tempo real.

Este trabalho, portanto, visa apresentar e comparar diversos modelos de aprendizado de máquina aplicados à classificação de mensagens eletrônicas, cujas amostras foram obtidas a partir de conjuntos de dados disponibilizados sob domínio público. Tais comparações serão feitas sob a luz de métricas — precisão, revocação, escore F_1 e tempos de treinamento e classificação — bem disseminadas no contexto acadêmico e relevantes ao tema (POWERS, 2011).

Por meio deste projeto, busca-se disseminar o entendimento de diversos modelos de aprendizado de máquina — conceito que chama cada vez mais atenção devido à sua intrínseca adaptabilidade — e evidenciar sua aplicação na solução de problemas práticos e de grande relevância nos dias de hoje.

[1]Conjunto de mecanismos cognitivos para filtragem de informação comunicada recebida: calibração de confiança, raciocínio, avaliação de argumentos e questionamento da plausibilidade de reivindicações.

1.2 Propostas existentes

Uma técnica frequentemente empregada por servidores de *e-mail* para lidar com *spam* é por meio do uso de listas negras (JUNG; SIT, 2004) (ZHANG et al., 2009).

O papel de uma DNSBL (do inglês *Domain Name System-based Blackhole List*, comumente conhecida como *blacklist* ou *block list*) é fornecer uma opinião a qualquer usuário que questionar se um determinado endereço IP (do inglês *Internet Protocol*) atende à política de aceitação de *e-mails* de entrada definida no servidor em questão. Tais listas podem ser consultadas no momento em que um novo *e-mail* é recebido pelo servidor, com o objetivo de manter mensagens maliciosas longe das caixas de entrada dos usuários.

Um exemplo de *blacklist* é a Spamhaus Block List (SBL) Advisory (The Spamhaus Project, 2018a), ilustrada na figura 4.

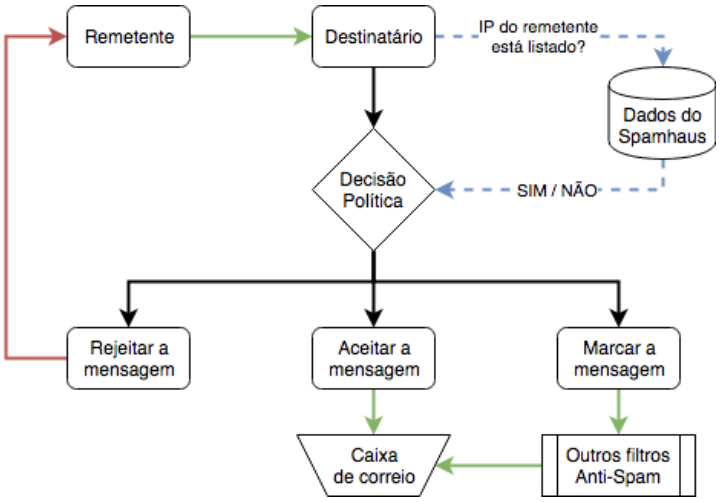


Figura 4: Funcionamento da SBL Advisory (The Spamhaus Project, 2018b).

Como uma etapa inicial adicional, podem ser consultadas listas brancas (*whitelists*), compostas por domínios ou endereços IP previamente aprovados e com permissão de entrega e que, portanto, são isentos de serem consultados em uma lista negra ou submetidos a posteriores verificações de presença de *spam*.

Uma outra abordagem ao problema de classificação de *e-mails* consiste em usar os dados contidos na própria mensagem (isto é, remetente, destinatário, assunto, cabeçalho, corpo, anexos etc.) para construir um sistema de decisão automático, inteligente e imparcial.

Para isto, é necessária a coleta de conjuntos de dados (ou seja, amostras reais de *e-mails*), sobre os quais um ou mais algoritmos de extração de características e aprendizado de máquina analisarão padrões inerentes a cada tipo de mensagem (legítima ou não) e tentarão derivar um conjunto de regras ou uma fronteira (e consequentemente uma função) de discernibilidade. O modelo obtido com o treinamento deste método poderá, então, ser usado para classificar novas mensagens com base no conhecimento obtido por meio da análise das mensagens anteriores.

Diversos autores propuseram o uso de modelos de aprendizado supervisionado, sobre *e-mails* previamente classificados da base de dados (ou seja, cada amostra já fora rotulada como *ham* ou *spam*). Utilizando diversos algoritmos, são construídos modelos capazes de inferir sobre a categoria de uma nova mensagem. Dentre as propostas existentes, pode-se citar o método probabilístico Naïve Bayes (ANDROUTSOPOULOS et al., 2000a) (MEYER; WHATELEY, 2004), redes neurais artificiais (CARPINTEIRO et al., 2006), máquina de vetores de suporte (DRUCKER et al., 1999) (ARAGÃO et al., 2016) (KUMARESAN; SARAVANAKUMAR; BALAMURUGAN, 2017) e *k-nearest neighbors* (BARIGOU; BELDJILALI; ATMANI, 2014).

Outros trabalhos adotaram modelos de aprendizado não-supervisionado, que também são capazes de apresentar taxas de precisão de classificação razoáveis, como na comparação entre máquina de vetores de suporte e mapas auto-organizáveis, apresentada em Berger e Merkl (2005).

1.3 Deficiências e necessidade de uma nova abordagem

Sistemas anti-*spam* baseados em listas negras e brancas apresentam diversas desvantagens, dentre as quais:

- Soluções pagas e de código proprietário: apesar das soluções proprietárias apresentarem altas taxas de captura de *spam* (SNYDER, 2015), muitas custam caro (da ordem de dezenas de milhares de dólares ao ano), comprometendo o orçamento de diversas instituições e/ou impedindo filtragem decente de *e-mails*. Além disso, licenças de código fechado impedem o estudo, personalização e extensão de funcionalidades por desenvolvedores independentes.
- Parcialidade questionável: diversos provedores oferecem serviços de consulta e remoção de endereços de *blacklists* mediante pagamento. Isto compromete toda a parcialidade e confiabilidade destes sistemas, já que virtualmente qualquer *spammer* (ou seja, pessoa ou organização que envia *spam*) pode deixar de sê-lo a qualquer momento.
- Fadadas à exaustão devido aos limites do IPv4/IPv6:
  - Se cada endereço IPv4 usar 4 B e seu nome de domínio em uma tabela DNS usar 8 B, seriam necessários $2^{32} * (4+8)\text{B} = 48\,\text{GiB}$ para armazenar estes dados. É um volume factível de armazenamento; entretanto, esta quantidade de endereços está próxima de se esgotar (ALI, 2012).
  - Se cada endereço IPv6 usar 16 B e seu nome de domínio em uma tabela DNS usar 8 B, seriam necessários $2^{128} * (16 + 8)\text{B} = 7.428 \times 10^{27}\,\text{TiB}$ para armazenar todos estes dados, o que é impraticável nos dias de hoje.
- Lentas (internet/latência): cada consulta a um servidor de *blacklist* está sujeita a uma latência, ou seja, o tempo que toda requisição leva para atingi-lo, ser processada e um resultado ser retornado ao cliente. Considerando um *round trip time* de 1 ms, um servidor de *e-mails* que processa 1 milhão de mensagens por dia gastaria 16 minutos e 40 segundos apenas para consultar o banco de endereços da *blacklist*. Caso seja feito um *cache* local dos dados, os problemas de armazenamento levantados no item anterior entram em cena.

1.4 Objetivos e contribuições da dissertação

O objetivo geral deste trabalho é aplicar modelos de aprendizado de máquina propostos por diversos autores à classificação de *e-mails*, a fim de obter modelos utilizáveis em sistemas anti-*spam* do mundo real.

Como objetivos e contribuições específicos, tem-se os seguintes pontos:

- pesquisar, estudar e comparar métodos de seleção de características e de redução de dimensionalidade, a fim de eleger aqueles que mais bem atendam às necessidades da proposta;
- elaborar um *framework* de *software* científico genérico, capaz de treinar diversos modelos de aprendizado de máquina e salvá-los para posterior reuso (e, em alguns casos, refinamento);
- expandir a funcionalidade do sistema anti-*spam* de código aberto OpenMaLBAS (do inglês *Open Machine-Learning-Based Anti-Spam*) (FERREIRA, 2018), fornecendo uma gama de métodos de classificação facilmente configuráveis e intercambiáveis;
- efetuar simulações em diversos conjuntos de dados e manter um histórico de resultados para fins de comparação;
- atingir desempenho satisfatório em termos de precisão, revocação, escore F_1 e tempos de treinamento e classificação, por meio da modificação de parâmetros de configuração de cada método;
- consolidar os resultados experimentais, comparando-os àqueles de outros trabalhos da comunidade acadêmica;
- apontar potenciais melhorias nos modelos propostos.

1.5 Estrutura da dissertação

A fundamentação teórica necessária para o desenvolvimento deste estudo está apresentada no capítulo 2. Nele são abordados conceitos de aprendizado de máquina, contemplando suas diferentes estratégias. Também apresenta conceitos de extração de características, filtragem de atributos, balanceamento de classes e elementos da teoria necessária sobre os algoritmos de classificação.

O capítulo 3 revisa, de forma cronológica, diversos trabalhos existentes sobre classificação de *e-mails*, abordando os modelos e métodos utilizados bem como os resultados alcançados em cada um.

Estão descritos no capítulo 4 os detalhes sobre os modelos avaliados neste estudo, desde o tratamento dos dados até a sua apresentação aos métodos de aprendizado de máquina, detalhando todos os seus componentes, inclusive ferramentas externas. Também descreve os conjuntos de dados utilizados. Por fim, esse capítulo aponta os resultados obtidos com a aplicação de cada modelo, inclusive sob a forma de *rankings* visando facilitar a interpretação dos resultados experimentais.

No capítulo 5, são apresentadas visualizações, análises e conclusões sobre os resultados obtidos, agrupados em três seções de forma a facilitar o entendimento de cada aspecto do trabalho.

Concluindo a dissertação, o capítulo 6 sumariza o problema e a proposta para resolvê-lo, bem como faz uma análise qualitativa dos resultados. Por fim, são apresentadas propostas para novos trabalhos.

Adicionalmente, é possível encontrar nos anexos uma cópia da licença sob a qual o código e os conjuntos de dados foram disponibilizados, e nos apêndices há tabelas com todos os resultados experimentais obtidos neste trabalho e detalhes do *hardware* e dos *softwares* utilizados.

2 Fundamentação Teórica

As seções a seguir visam explicar os principais conceitos necessários ao entendimento da proposta deste trabalho. Primeiramente, são introduzidos o conceito e os tipos de aprendizado de máquina, base para toda a aquisição de conhecimento e construção de sistemas computacionais inteligentes. Também visita-se o conceito de classificação de texto, essencial à categorização de quaisquer exemplos de dados multimídia compostos por caracteres (em vez de valores numéricos reais), como por exemplo os *e-mails*, objeto de estudo desta proposta.

Em seguida, as três técnicas de seleção de características adotadas no trabalho são apresentadas, bem como o método de redução de dimensionalidade e de balanceamento de classes aplicado aos conjuntos de dados. A finalidade de tais técnicas é refinar as amostras de treinamento, removendo informações redundantes e garantindo uma base de conhecimento apropriada à construção dos modelos de classificadores.

Na sequência, cada técnica usada para classificação das mensagens é categorizada quanto à sua natureza (métodos probabilísticos, redes neurais artificiais, árvores de decisão e máquina de vetores de suporte) e explicada. Finalmente as métricas para avaliação dos resultados são apresentadas e definidas matematicamente.

2.1 Aprendizado de máquina

O aprendizado é um processo que acompanha o homem desde o início de sua existência. Quando representados de forma coerente e inteligível, estímulos sensoriais desencadeiam todo um de processo de organização, associação e geração de conhecimento. O artefato deste processo é a capacidade natural de inferência, conferindo ao homem as capacidade de raciocínio, discernimento e tomada de decisão, mesmo diante de temas sobre os quais nunca havia lidado.

Desde o final da década de 1950 (SAMUEL, 1959), vem-se tentando imitar o complexo processo de aprendizado em ambientes computacionais, a fim de conceber sistemas inteligentes e que possam auxiliar (ou até mesmo substituir) o homem em situações que demandam tomadas sistemáticas de decisão. Esta área do conhecimento recebeu o nome de aprendizado de máquina e é um assunto cada vez mais em voga na comunidade acadêmica e no meio científico.

As tarefas de aprendizagem de máquinas podem ser classificadas em três categorias, dependendo da presença e do tipo de *feedback* de aprendizado disponível para um sistema de aprendizagem (RUSSELL; NORVIG, 2013):

- Aprendizagem não-supervisionada: o agente aprende padrões na entrada, embora não seja fornecido nenhum *feedback* explícito. A tarefa mais comum de aprendizagem não supervisionada é o agrupamento: a detecção de grupos de exemplos de entrada potencialmente úteis.

- Aprendizagem por reforço: o agente aprende a partir de uma série de reforços — recompensas ou punições. Cabe ao agente decidir qual das ações anteriores ao reforço foram as maiores responsáveis por isso.

- Aprendizagem supervisionada: o agente observa exemplos de pares de entrada e saída, e aprende uma função que faz o mapeamento da entrada para a saída.

A figura 5 ilustra um processo de aprendizado não-supervisionado, onde faz-se o agrupamento dos vetores de entrada com base em alguma medida de similaridade entre suas características. Para um novo dado, a saída do modelo preditivo pode ser uma probabilidade, uma indicação do grupo mais provável ou uma representação melhorada deste dado, com base no conhecimento adquirido durante o treinamento.

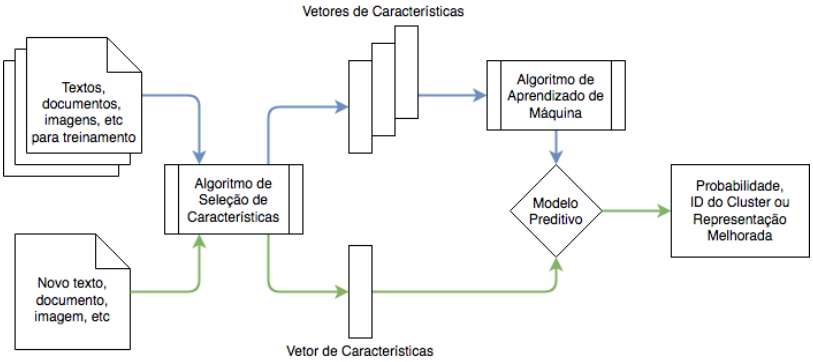


Figura 5: Aprendizado não-supervisionado.

A figura 6 ilustra um processo de aprendizado supervisionado, onde cada par vetor-rótulo de entrada é analisado de forma a obter uma função capaz de mapear um novo dado de entrada a um dos rótulos conhecidos pelo algoritmo. Para um novo dado, portanto, a saída do modelo preditivo é um rótulo que indica a qual classe este dado provavelmente pertence, com base no conhecimento adquirido durante o treinamento.

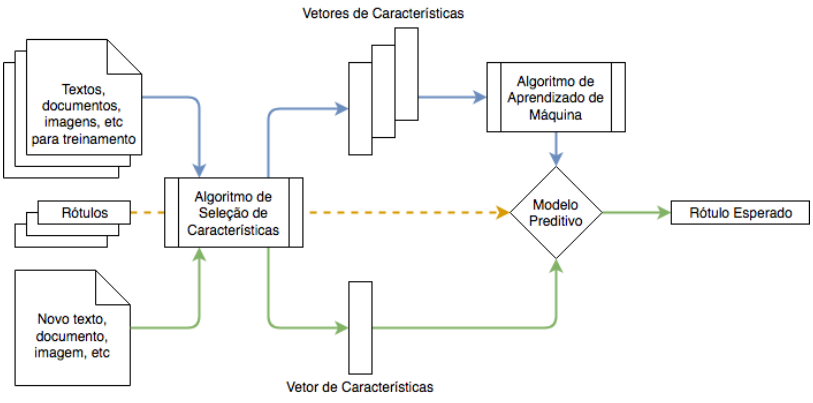


Figura 6: Aprendizado supervisionado.

2.2 Categorização de texto

A categorização de texto é tarefa de atribuir categorias predefinidas a documentos de texto livre. Ela pode ser útil nos mais diversos contextos e aplicações do mundo real. A tabela 1 mostra como este conceito pode ser aplicado a certos tipos de texto (YANG; JOACHIMS, 2008):

Tabela 1: Exemplos de tipos de texto e sugestões de classificação.

Tipo de texto	Exemplos de categorização
Notícias	Assunto, tópico ou código geográfico
Trabalhos acadêmicos	Domínios técnicos e subdomínios
Prontuários médicos	Tipo de doença, procedimentos cirúrgicos e medicação
E-mail	*Ham*, *spam*, pessoais ou de negócios
Redes sociais	Postagens casuais, difamatórias e *fake news*

Em vez de manualmente classificar documentos ou criar regras de classificação automática, é possível utilizar métodos de aprendizado de máquina para aprender regras de classificação automática com base em documentos de treinamento que foram previamente categorizados por humanos.

Para isto, é necessário primeiramente representar estes documentos (que, no contexto deste trabalho, consistem em mensagens de *e-mail*) em uma forma passível de processamento computacional. Esta representação, que consiste em um conjunto de características do documento, pode ser simplificada por meio de métodos de seleção de características. Alguns destes métodos são apresentadas na seção 2.3. Um estudo mais aprofundado pode ser encontrado em Forman e George (2003).

Em seguida, entram em cena os modelos de aprendizado de máquina que, por meio de treinamento (usando vetores de características extraídas de amostras de texto), são capazes de classificar textos, *e-mails* etc. Vários destes modelos são apresentados na seção 2.6. Um levantamento mais abrangente pode ser encontrado em Aggarwal e Zhai (2012).

2.3 Seleção de características

Os problemas de categorização de texto normalmente apresentam espaços de característica de alta dimensionalidade, já que são compostos por todos os termos únicos (palavras, por exemplo) que ocorrem nos documentos do conjunto. Este grande volume de dados de entrada, que muitas vezes inclui características redundantes ou irrelevantes, frequentemente atinge a casa de centenas ou milhares de características, podendo dificultar (ou até mesmo impedir) o treinamento e a geração de modelos com certos algoritmos de aprendizado de máquina.

Diante deste problema, mostra-se necessário efetuar uma redução do conjunto de características, de modo que os algoritmos não gastem quantidades impraticáveis de tempo e recursos (armazenamento e processamento) computacionais. Além disso, deseja-se escolher um subconjunto de características que retenha a maior quantidade possível de informação do conjunto original, de forma a não prejudicar a capacidade de generalização e precisão de categorização do modelo.

Em Yang e Pedersen (1997), são revisitados e comparados diversos métodos estatísticos de seleção de características, especificamente aplicados à categorização de textos, tais como "seleção de termos baseada na frequência no documento, ganho de informação, informação mútua, teste χ^2 e força do termo". É importante ressaltar que os termos não são necessariamente palavras, mas qualquer elemento presente no *e-mail* após o pré-processamento (ex: *tokens* para caracterizar *tags* HTML (do inglês *HyperText Markup Language*), números, pontuação, *links* etc.).

Na literatura, também é possível encontrar métodos de extração de características que visam criar conjuntos de novas características definidas em função das originais. Apesar de tais métodos fugirem ao escopo deste trabalho, podem ser citadas as abordagens baseadas em análise semântica latente (DEERWESTER et al., 1990) (HOFMANN, 1999), mapeamento isométrico de características (TENENBAUM; SILVA; LANGFORD, 2000) e incorporação estocástica de vizinhos (HINTON; ROWEIS, 2002) (MAATEN; HINTON, 2008).

A seguir, serão explicados os três métodos de seleção de características — estatística χ^2, frequência no documento e informação mútua — utilizados neste trabalho.

2.3.1 Estatística Qui-quadrado (CHI2)

A estatística χ^2 (ou simplesmente CHI2) mede a dependência entre um termo t e uma classe c em particular. É definida pela equação (2.1).

$$CHI2(t,c) = \frac{n \cdot F(t)^2 \cdot (p_c(t) - P_c)^2}{F(t) \cdot (1 - F(t)) \cdot P_c \cdot (1 - P_c)} \tag{2.1}$$

Na qual:

- n é o número total de documentos na coleção;
- $p_c(t)$ é a probabilidade condicional c para documentos que contenham t;
- P_c é a fração global de documentos que contém a classe c;
- $F(t)$ é a fração global de documentos que contem o termo t.

Uma característica deste método é que os valores $CHI2(t,c)$ são normalizados e, portanto, são comparáveis entre termos de uma mesma categoria.

2.3.2 Frequência no Documento (FD)

É um método estatístico que mede a frequência que um termo t ocorre em uma classe c, em particular. É definido pela equação (2.2).

$$FD(t,c) = \frac{C(t,c)}{\sum_{t_i \in T(c)} C(t_i,c)} \tag{2.2}$$

Na qual:

- o numerador é a quantidade de ocorrências de t nos documentos da classe c;
- o denominador é o somatório das quantidades de ocorrências de todos os termos nos documentos da classe c.

2.3.3 Informação Mútua (MI)

Derivada da teoria da informação (COVER; THOMAS, 1991), esta medida consiste na quantidade de informação que um termo agrega em relação a uma dada classe. A informação mútua pontual $M(t, c)$ entre o termo t e a classe c é baseada no nível de coocorrência entre a classe c e o termo t, dada pela equação (2.3).

$$MI(t, c) = \log\left(\frac{F(t) \cdot p_c(t)}{F(t) \cdot P_c}\right) = \log\left(\frac{p_c(t)}{P_c}\right) \tag{2.3}$$

Na qual:

- $F(t) \cdot P_c$ é a coocorrência esperada da classe x e termo t com base na independência mútua;
- $F(t) \cdot p_c(t)$ é a coocorrência verdadeira (na prática, esse valor pode ser muito maior ou menor que o esperado, dependendo do nível de correlação entre a classe x e o termo t).

Claramente, o termo t é positivamente correlacionado à classe c quando $MI(t, c) > 0$, e negativamente correlacionada quando $MI(t, c) < 0$.

Note que as equações (2.1), (2.2) e (2.3) são computadas para uma única classe. Para criar um valor que abranja todas as classes, podem ser usados os valores médios ou máximos, como nas equações (2.4), (2.5), (2.6), (2.7), (2.8) e (2.9):

$$CHI2_{médio}(t) = \sum_{c \in C} P_i \cdot \chi^2(t, c) \tag{2.4}$$

$$CHI2_{máximo}(t) = \max_{c \in C} \chi^2(t, c) \tag{2.5}$$

$$FD_{médio}(t) = \frac{1}{|C|} \sum_{c \in C} FD(t, c) \tag{2.6}$$

$$FD_{máximo}(t) = \max_{c \in C} FD(t, c) \tag{2.7}$$

$$MI_{médio}(t) = \sum_{c \in C} P_i \cdot MI(t, c) \tag{2.8}$$

$$MI_{máximo}(t) = \max_{c \in C} MI(t, c) \tag{2.9}$$

2.4 Redução de dimensionalidade

Um problema central no aprendizado de máquina é identificar um conjunto representativo de atributos (características) para construir um modelo de classificação para uma tarefa específica. Em modelos preditivos, atributos em excesso podem confundir o classificador, tornando a redução de dimensionalidade uma etapa importante no processo de aprendizado.

A "maldição da dimensionalidade" refere-se a vários fenômenos que surgem ao analisar e organizar dados em espaços de alta dimensão (muitas vezes com centenas ou milhares de dimensões) que não ocorrem em espaços de baixa dimensão, como o espaço físico tridimensional da experiência cotidiana. (DONOHO, 2000)

No aprendizado de máquina, muitos problemas exigem grandes quantidades de dados de treinamento para garantir que existam várias amostras com valores distintos. Com um número fixo de amostras de treinamento, o poder preditivo de um classificador ou regressor inicialmente aumenta à medida que o número de dimensões (características) usados é aumentado, mas depois decai. Isto é conhecido como "fenômeno de Hughes" (HUGHES, 1968), e é ilustrado na figura 7.

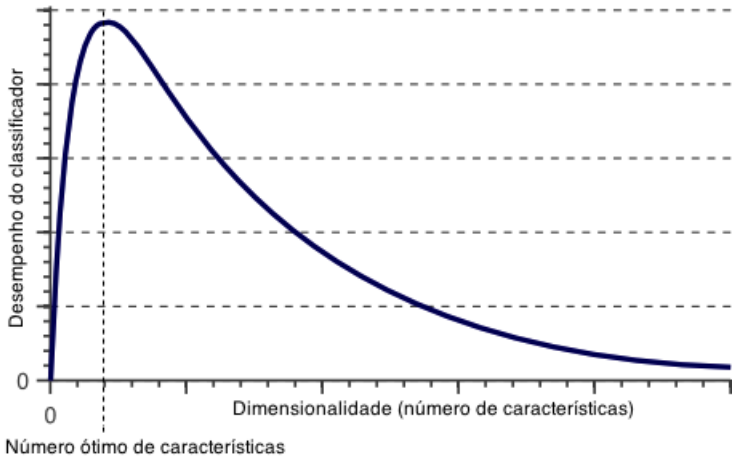


Figura 7: Maldição da dimensionalidade (traduzido) (SPRUYT, 2014).

Diversas técnicas de redução de dimensionalidade foram propostas, tais como:

- Análise de componentes principais (PCA, do inglês *Principal Component Analysis*): utiliza a ortogonalização de vetores para converter um conjunto de variáveis possivelmente correlacionadas num conjunto de variáveis linearmente não correlacionadas (chamadas de componentes principais). (F.R.S., 1901)
- Consistência de subconjunto (LVF, do inglês *Las Vegas Filter*): baseia-se em um algoritmo de busca aleatório que demonstra alta robustez (em termos de atributos redundantes ou correlações de ordem alta) quando utilizado com conjuntos de dimensionalidade alta. (LIU; SETIONO, 1996)
- Correlação de subconjunto (CFS, do inglês *Correlation-based Feature Selection*): baseia-se na hipótese central de que bons conjuntos de características possuem características altamente correlacionadas com a classe, ainda que não correlacionadas entre si. (HALL, 2000)
- Máquina de vetores de suporte (SVM-RFE, do inglês *Support Vector Machine Recursive Feature Elimination*): faz uso de SVMs com eliminação recursiva de características para efetuar redução de dimensionalidade, descartando atributos redundantes e produzindo subconjuntos compactos e, ao mesmo tempo, com grande capacidade de predição. (GUYON et al., 2002)
- Filtro de silhueta simplificada (SSF, do inglês *Simplified Silhouette Filter*): consiste em agrupar atributos em subconjuntos de cardinalidade estimada automaticamente, podendo considerar correlações entre características e classes. (COVÕES; HRUSCHKA, 2011)

Neste trabalho, foi adotada a técnica CFS devido a sua grande capacidade de redução de dimensionalidade em uma variedade de conjuntos de dados (com características tanto discretas quanto contínuas), ao mesmo tempo que mantém ou aumenta o desempenho dos algoritmos de aprendizado. (HALL, 2000)

O algoritmo CFS primeiro calcula uma matriz de correlações entre características-classes e características-características a partir do conjunto de treinamento e, em seguida, realiza uma busca pelo subconjunto com maior nota. Para esta última tarefa, foi utilizado o algoritmo evolucionário multiobjetivo ENORA (do inglês *Evolutionary NOn-dominated Radial slots based Algorithm*) (JIMENEZ et al., 2002).

2.5 Balanceamento de classes

A finalidade desta etapa é assegurar que haja a mesma quantidade de amostras de *hams* e *spams*. Isto porque um conjunto de dados desbalanceado (com quantidades diferentes de instâncias de cada classe) pode fazer com que o método de aprendizado de máquina enfatize mais uma classe do que outra, obtendo resultados enviesados. Para resolver esta situação, pode-se selecionar aleatoriamente mensagens no conjunto de dados de menor cardinalidade e replicá-las até que os conjuntos possuam o mesmo tamanho. Considere o conjunto de dados da tabela 2, onde k é o índice da amostra, x_i^k é o valor da i-ésima característica da k-ésima amostra e y^k é a classe da k-ésima amostra.

Tabela 2: Conjunto de dados antes do balanceamento.

k	x_0^k	x_1^k	x_2^k	x_3^k	y^k
1	0,24	0,59	0,81	0,04	*ham*
2	0,66	0,24	0,86	-0,11	*ham*
3	0,48	0,15	0,76	0,68	*ham*
4	0,53	0,39	0,34	0,14	*ham*
5	0,08	0,56	0,01	0,55	*ham*
6	1,14	0,25	0,81	1,13	*spam*
7	-0,44	-0,54	-0,10	0,78	*spam*
8	0,36	0,59	0,30	0,91	*spam*

Neste caso, a classe *ham* possui cinco amostras, enquanto a *spam* possui apenas três. Para resolver esta situação, escolhem-se aleatoriamente $5 - 3 = 2$ amostras de *spam*, que são replicadas no conjunto de dados visando balancear as cardinalidades (vide tabela 3). Assim, as classes passam a quantidades iguais de amostras, equalizando as influências de cada uma perante o algoritmo de aprendizado.

Tabela 3: Amostras replicadas a fim de balancear o conjunto de dados.

k	x_0^k	x_1^k	x_2^k	x_3^k	y^k
9	-0,44	-0,54	-0,10	0,78	*spam*
10	1,14	0,25	0,81	1,13	*spam*

2.6 Algoritmos de classificação

A classificação consiste em, utilizando um conjunto de dados com amostras rotuladas e um algoritmo de aprendizado de máquina, criar um modelo capaz de atribuir classes a dados inéditos. Existem diversos modelos de classificação, e uma grande variedade é estudada de forma abrangente em Bishop (2006). No escopo deste trabalho, fazem-se relevantes os métodos discriminados na listagem a seguir, divididos em quatro categorias:

- Métodos probabilísticos (subseção 2.6.1): consiste em classificadores capazes de prever, dada uma observação de entrada, uma distribuição de probabilidade em um conjunto de classes, em vez de apenas produzir a classe mais provável a que a observação deveria pertencer.
- Redes neurais artificiais (subseção 2.6.2): são modelos computacionais inspirados pelo sistema nervoso central de um animal, capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões.
- Árvores de decisão (subseção 2.6.3): é uma ferramenta de suporte à decisão que usa um grafo ou modelo de decisões em árvore e suas possíveis consequências; são comumente usados na pesquisa de operações para ajudar a identificar uma estratégia com maior probabilidade de atingir um objetivo, sendo também uma ferramenta popular na aprendizagem de máquinas.
- Máquina de vetores de suporte (subseção 2.6.4): é um classificador discriminativo formalmente definido por um hiperplano separador; em outras palavras, com dados de treinamento rotulados (aprendizagem supervisionada), o algoritmo produz um ótimo hiperplano que categoriza novos exemplos.

Algoritmos que não foram abordados neste trabalho mas que merecem destaque incluem regressão logística, modelos ocultos de Markov, aprendizado por quantização vetorial (KOHONEN, 1995) e *gradient boosting* (FRIEDMAN, 2001).

2.6.1 Métodos Probabilísticos

2.6.1.1 Naïve Bayes (NB)

Naïve Bayes é um algoritmo para classificação conhecido tanto por sua simplicidade teórica e facilidade de implementação, quanto também pela eficácia (especialmente na categorização de textos), já que é capaz de construir modelos precisos e fazer previsões rapidamente. Pode ser utilizado em diversos cenários, tais como filtragem de *spam*, análise sentimental, categorização de notícias, diagnósticos médicos, previsão do tempo e reconhecimento facial e de caracteres.

O algoritmo aprende a probabilidade de um objeto com determinadas características pertencer a uma determinada classe ou categoria; por esta razão, é tido como um "classificador probabilístico". Além disso, é dito "bayesiano" por ter como base o teorema de Bayes e "ingênuo" por supor que a ocorrência de uma determinada característica independe da ocorrência e/ou influência de outras características.

Proposto pelo matemático Thomas Bayes em 1763 e aprimorado pelo matemático, astrônomo e físico Pierre-Simon Laplace em 1812, o teorema (dado pela equação (2.10)) descreve a probabilidade de ocorrência de um evento com base no conhecimento prévio de condições que possam estar relacionadas ao evento.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{2.10}$$

em que:

- A e B são eventos e $P(B) \neq 0$;
- $P(A|B)$ é a probabilidade da ocorrência do evento A dado que B ocorreu;
- $P(A)$ e $P(B)$ são as probabilidades de ocorrência de A e B, respectivamente;
- $P(B|A)$ é a probabilidade da ocorrência do evento B dado que A ocorreu.

Dado um vetor de dados $\vec{x} = \{x_1, \ldots, x_n\}$ com n características, o classificador prediz a classe C_k para $\vec{x}$ de acordo com a equação (2.11) (MCGONAGLE, 2018):

$$p(C_k|\vec{x}) = p(C_k|x_1, \ldots, x_n) \text{ para } k = 1, \ldots, K \tag{2.11}$$

Usando o teorema de Bayes, a equação (2.11) pode ser fatorada como:

$$p(C_k|\vec{x}) = \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} = \frac{p(x_1, \ldots, x_n|C_k)p(C_k)}{p(x_1, \ldots, x_n)} \tag{2.12}$$

Usando a regra da cadeia, o fator $p(x_1, \ldots, x_n|C_k)$ no numerador da equação (2.12) pode ser decomposto ainda mais em:

$$p(x_1, \ldots, x_n|C_k) = p(x_1|x_2 \ldots x_n, C_k)p(x_2|x_3 \ldots x_n, C_k) \ldots p(x_{n-1}|x_n, C_k)p(x_n|C_k) \tag{2.13}$$

Aplicando a suposição ingênua de independência condicional (ou seja, assumindo que a característica x_i é independente da característica x_j para $i \neq j$ dada a classe C_k) à equação (2.13), tem-se:

$$p(x_1, \ldots, x_n|C_k) = p(x_i|C_k) \Rightarrow p(x_1, \ldots, x_n|C_k) = \prod_{i=1}^{n} p(x_i|C_k) \tag{2.14}$$

Desta forma,

$$\begin{aligned} p(C_k|x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)p(x_1, \ldots, x_n|C_k) \\ &\propto p(C_k)p(x_1|C_k)p(x_2|C_k) \ldots p(x_x|C_k) \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \end{aligned} \tag{2.15}$$

O classificador NB combina o modelo descrito pelas equações anteriores com uma regra de decisão, como a estimativa de probabilidade máxima *a posteriori*, que consiste na escolha da hipótese mais provável. O classificador correspondente é a função que atribui a classe $\hat{C} = C_k$ para algum k como descrito na equação (2.16):

$$\hat{C} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k). \tag{2.16}$$

Além do NB, há outras abordagens que fazem uso da suposição de independência. No trabalho de McCallum e Nigam (1998), por exemplo, são comparados os modelos multivariado de Bernoulli e multinomial aplicados à classificação de texto.

2.6.1.2 Averaged One-Dependence Estimators (AODE)

Conforme explicado na seção 2.6.1.1, o classificador NB baseia-se na suposição ingênua de independência entre atributos. Entretanto, em conjuntos obtidos de contextos em que os atributos são fortemente relacionados, tal premissa pode acarretar em altas taxas de erro quando da classificação de amostras. Neste sentido, Sahami (1996) introduz a noção de estimadores de x dependências, por meio dos quais a probabilidade do valor de cada atributo é condicionada pela classe e, no máximo, outros x atributos, induzindo um classificador "menos ingênuo" que o NB.

Obviamente, a introdução de dependências entre atributos aumenta exponencialmente a quantidade de poder computacional necessária para estimar as probabilidades condicionais. Entretanto, estes dados podem ser pré-computados uma única vez (ou seja, antes do treinamento propriamente dito), minimizando assim o *overhead* causado por essa suposição menos ingênua.

Neste trabalho, será adotado $x = 1$, ou seja, o foco será em "estimadores de uma dependência" (AODE). Neste cenário, um problema rapidamente vem à tona: um classificador deste tipo requer que cada atributo dependa de outro atributo, e tal atributo específico deve ser necessariamente selecionado. A solução para este impasse consiste em considerar todos os classificadores de uma dependência e agregar as previsões de todos eles, desde que estejam qualificados para tal.

Sucintamente, um classificador é qualificado para contribuir para a previsão final somente se possuir um único atributo que seja pai de todos os outros e se houver no mínimo m amostras no conjunto de treinamento com o atributo pai preenchido, ou seja, com um valor não-nulo. Este último critério visa incluir modelos para os quais as estimativas básicas de probabilidades são imprecisas. Além disso, no mesmo trabalho, adota-se $m = 30$, "por ser uma quantidade mínima de amostras amplamente utilizada para propósitos de inferência estatística" (WEBB; BOUGHTON; WANG, 2005).

As figuras 8a e 8b ilustram o funcionamento do NB e do AODE, respectivamente. Na primeira, nota-se que as probabilidades de classe $P(c)$ dependem diretamente dos atributos $x_1, x_2, \ldots, x_n$, e nenhuma probabilidade entre eles é considerada. Já na segunda, $P(c)$ está condicionada aos atributos que, por sua vez, levam em consideração as probabilidades conjuntas em relação a um único outro atributo, caracterizando assim um estimador de uma dependência.

(a) Funcionamento do NB. (b) Funcionamento do AODE.

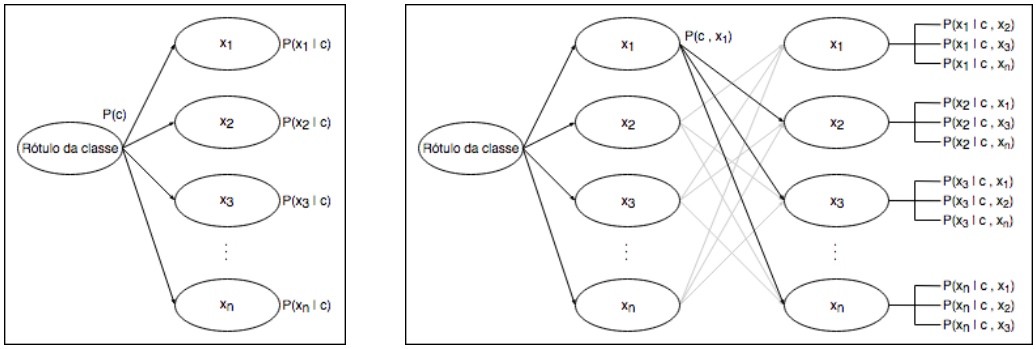


Figura 8: Comparação entre dependências dos classificadores NB e AODE.

O funcionamento do AODE consiste, portanto, em estimar as probabilidades de uma amostra $\vec{x}$ com valores de características $x_1, x_2, \ldots, x_n$ pertencer a cada classe y_i e, em seguida, eleger a classe final $\hat{y}$ como aquela que maximiza a equação (2.17). Nela, o termo $F(x_i)$ representa a quantidade de exemplos que possuem o valor de atributo x_i, e é usado para assegurar o limite m de amostras necessárias para um dado classificador poder contribuir para a estimativa de probabilidade (WEBB; BOUGHTON; WANG, 2005).

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left(\sum_{i:1 \leq i \leq n \wedge F(x_i) \geq m} P(y, x_i) \prod_{j=1}^{n} P(x_j \mid y, x_i) \right) \tag{2.17}$$

É importante destacar que, se não existir um classificador que satisfaça a condição de qualificação, a consideração de uma dependência não será levada em conta e, portanto, o método AODE comportar-se-á como o NB.

2.6.2 Redes Neurais Artificiais

As redes neurais artificiais são modelos matemáticos compostos por unidades discretas de processamento, cujo funcionamento baseia-se no de um neurônio biológico. Quando dispostos em uma ou mais camadas neurais associadas por meio de conexões direcionadas, tais elementos passam a caracterizar uma rede neural, estruturas frequentemente comparadas ao cérebro humano.

A figura 9 apresenta um modelo simples do neurônio artificial (MCCULLOCH; PITTS, 1943). Se a combinação linear de suas entradas exceder um limiar pré determinado, o neurônio produzirá um potencial excitatório em sua saída; caso contrário, o neurônio produzirá um potencial inibitório.

Cada unidade de uma rede neural possui uma entrada fictícia $a_0 = 1$ como um peso $w_{0,j}$ associado a ela. Uma ligação da unidade i para a unidade j serve para propagar a ativação a_i de i para j, e o peso desta ligação (chamado de peso sináptico, já que as "sinapses" são as regiões onde ocorre a transmissão de impulsos nervosos em um neurônio biológico), representado por $w_{i,j}$, determina a força e o sinal da conexão. Cada unidade j primeiro calcula uma soma ponderada de suas entradas e, em seguida, aplica uma função de ativação g a essa soma para obter a saída $a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$. (RUSSELL; NORVIG, 2013)

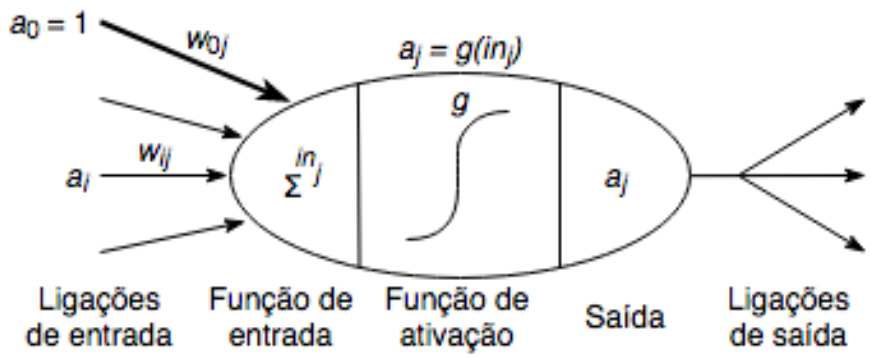
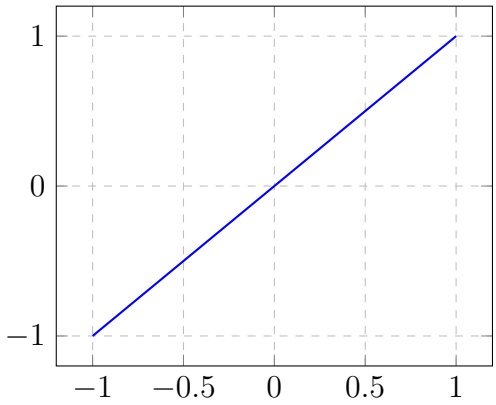


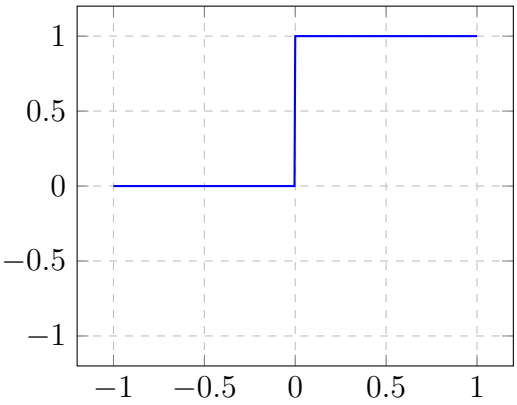
Figura 9: Neurônio artificial de McCulloch–Pitts (MCP) (traduzido).

A função g é tipicamente uma das listadas nos gráficos da figura 10.

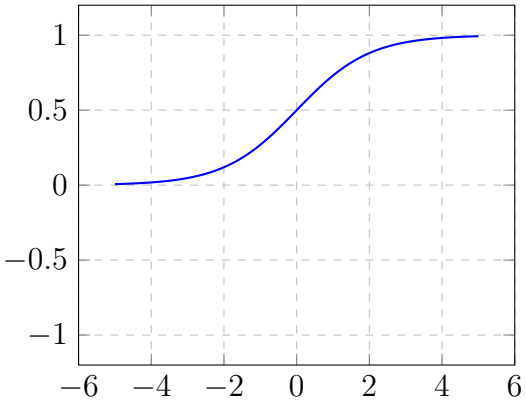
(a) Função Identidade.



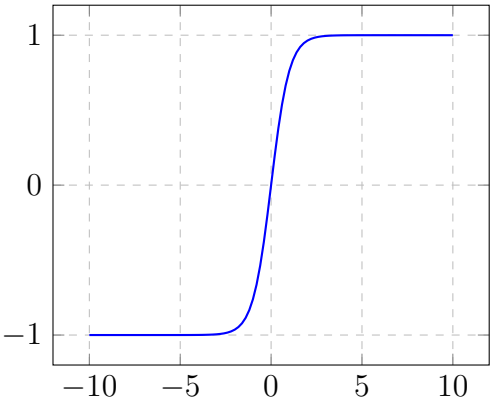
(b) Função Degrau Binário.



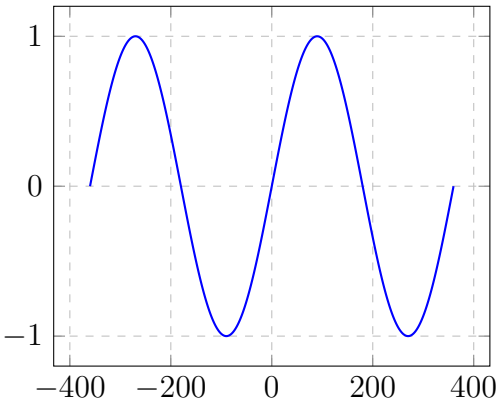
(c) Função Sigmoide.



(d) Função Tangente Hiperbólica.



(e) Função Senoide (ou Sinusoide).



(f) Função Gaussiana.

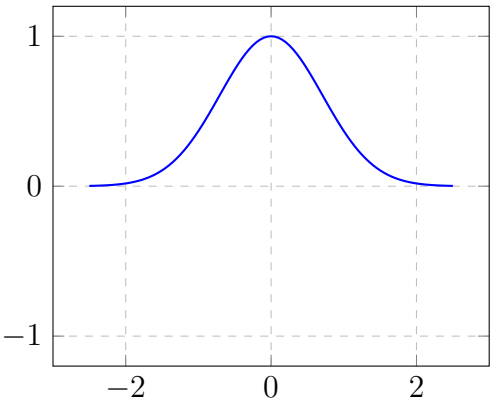


Figura 10: Exemplos de funções de ativação.

2.6.2.1 Single Layer Perceptron (SLP)

Uma rede neural com todas as entradas conectadas diretamente às saídas é dita simples, de camada única ou rede Perceptron. A figura 11 mostra um exemplo deste tipo de rede com duas entradas (x_1 e x_2) e uma saída (y).

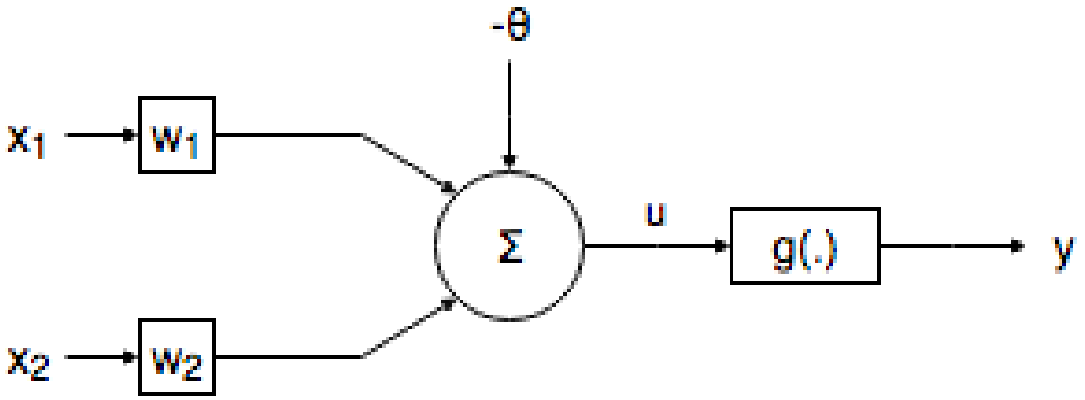


Figura 11: Rede Perceptron de camada única.

A capacidade de cálculo do Perceptron está limitada pela separabilidade linear (vide figuras 12a e 12b) das classes. Sua incapacidade em lidar com classes distribuídas de tal maneira que seja impossível criar um hiperplano para separá-las (vide figura 12c) foi um revés significativo para as redes neurais na década de 1960. Nestes casos, faz-se necessário utilizar abordagens mais sofisticadas, como uma rede de função de base radial, que é discutida adiante na seção 2.6.2.2.

(a) Função $x_1 \wedge x_2$. (b) Função $x_1 \vee x_2$. (c) Função $x_1 \oplus x_2$.

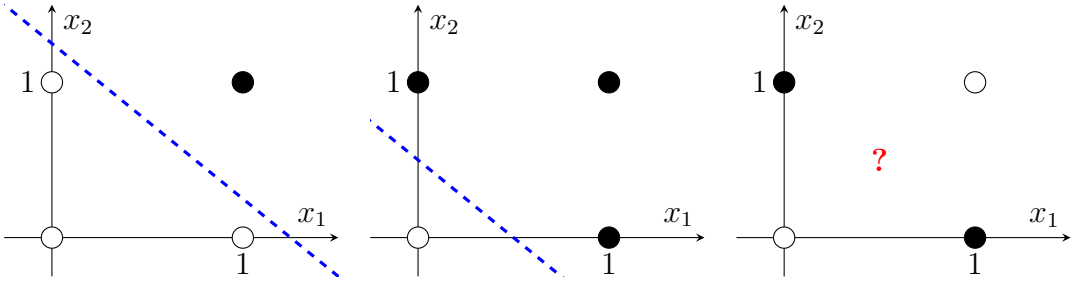


Figura 12: Funções binárias e separabilidade linear.

O treinamento de uma rede Perceptron consiste em ajustar os valores dos pesos e limiar, e baseia-se na Teoria Hebbiana (HEBB, 1949). Resumidamente, se a saída produzida pelo Perceptron difere da desejada, seus pesos sinápticos e limiares serão então ajustados proporcionalmente aos valores dos sinais de entrada. Este processo é repetido para todas as amostras de treinamento até que a saída produzida pela rede seja similar à saída desejada de cada amostra (ou até que outro critério de parada, por exemplo quantidade de épocas de treinamento, seja alcançado).

Em termos matemáticos, as regras de ajuste dos pesos sinápticos $\vec{w}$ e do limiar θ do neurônio podem ser expressas pela equação (2.18):

$$\vec{w}_i^{atual} = \vec{w}_i^{anterior} + \eta \cdot \Big(d^{(k)} - y\Big) \cdot \vec{x}^{(k)} \tag{2.18}$$

Na qual:

- $\vec{w} = \begin{bmatrix}\theta & w_1 & w_2 & \dots & w_n\end{bmatrix}^\top$ é o vetor de pesos sinápticos;
- $\vec{x}^{(k)} = \begin{bmatrix}-1 & x_1^{(k)} & x_2^{(k)} & \dots & x_n^{(k)}\end{bmatrix}^\top$ é a k-ésima amostra de treinamento;
- $d^{(k)}$ é o valor desejado para a k-ésima amostra de treinamento;
- y é o valor de saída produzido pelo Perceptron;
- η é uma constante que define a taxa de aprendizagem da rede.

A taxa de aprendizagem η "exprime o quão rápido o processo de treinamento da rede será conduzido rumo à convergência (estabilização) e, normalmente, são adotados valores entre 0 e 1" (SILVA; SPATTI; FLAUZINO, 2010).

Outra técnica frequentemente adotada para treinar este tipo de rede baseia-se no algoritmo LMS (do inglês *Least Mean Squares*) (WIDROW; HOFF, 1960). Esta abordagem visa ajustar os pesos e limiar do neurônio de forma a minimizar o erro da rede (ou seja, as diferenças entre as saídas desejadas e as respostas do combinador linear) na classificação das amostras de treinamento. O erro da rede é normalmente medido usando a função EQM (Erro Quadrático Médio), dada pela equação (2.19).

$$E(\vec{w}) = \frac{1}{2}\sum_{k=1}^{p}\Big(d^{(k)} - (\vec{w}^\top \cdot \vec{x}^{(k)} - \theta)\Big)^2 \tag{2.19}$$

2.6.2.2 Radial Basis Function Network (RBF)

As redes Perceptron de múltiplas camadas (WHITE; ROSENBLATT, 1963) são caracterizadas pela presença de pelo menos uma camada intermediária (escondida) de neurônios, situada entre as camadas de entrada e de saída. Este tipo de rede pode ser aplicado em diversos tipos de problemas, tais como aproximação universal de funções, reconhecimento de padrões, identificação e controle de processos, previsão de séries temporais e otimização de sistemas (SILVA; SPATTI; FLAUZINO, 2010).

As redes denominadas de funções de base radial (BROOMHEAD; LOWE, 1988), convencionalmente conhecidas como RBF (do inglês *Radial Basis Function*), assim como as Perceptron de camadas múltiplas, pertence à arquitetura *feedforward*, já que o fluxo de informações na sua estrutura se inicia na camada de entrada, percorre a camada intermediaria única (neurônios com funções de ativação $g_1^{(1)}, g_2^{(1)}, \ldots, g_{n_1}^{(1)}$ do tipo gaussiana) e finaliza na camada neural de saída (neurônios com funções de ativação $g_1^{(2)}, \ldots, g_{n_2}^{(2)}$ lineares). A figura 13 ilustra a topologia deste tipo de rede.

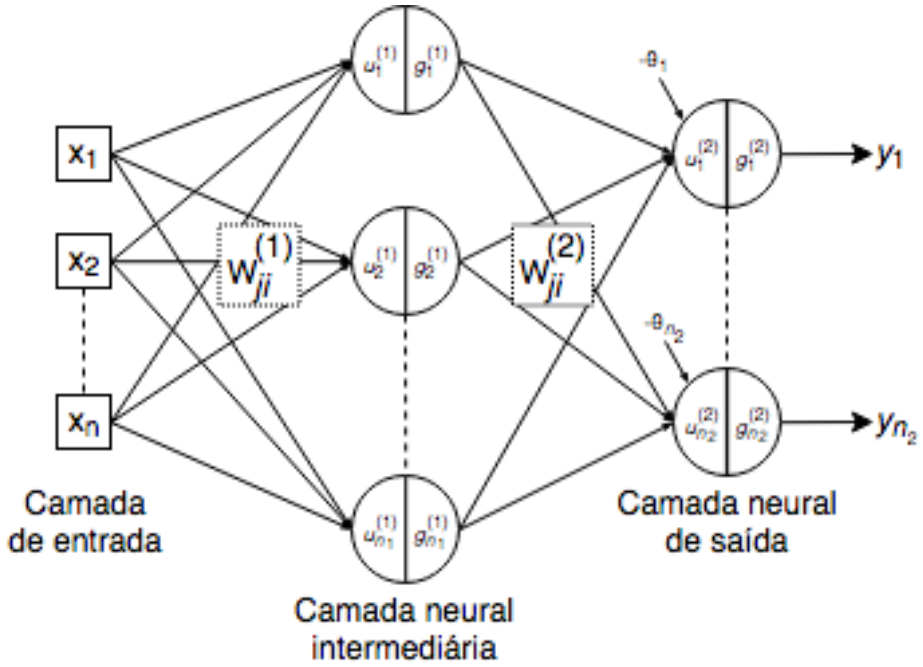


Figura 13: Rede de função de base radial.

A estratégia de treinamento utilizada para o ajuste dos pesos de suas duas únicas camadas neurais consiste de dois estágios:

1. Ajustes dos pesos dos neurônios da camada intermediária usando um método de aprendizagem não-supervisionado (ou seja, que só utiliza as características dos dados de entrada) para posicionar as funções de base radial.
2. Ajustes dos pesos dos neurônios da camada de saída, utilizando um critério de aprendizagem baseado em retropropagação do erro, frequentemente referido como regra delta generalizada (RUMELHART; HINTON; WILLIAMS, 1985).

Além da estratégia de treinamento recém citada, cujos detalhes são apresentados ainda nesta seção, foi proposta na literatura uma versão totalmente supervisionada deste tipo de rede, nomeada GRBF (do inglês *Generalized Radial Basis Functions*) (POGGIO; GIROSI, 1989) (WETTSCHERECK; DIETTERICH, 1992).

Etapa 1: treinamento não-supervisionado Como dito anteriormente, este estágio consiste em ajustar os pesos dos neurônios da camada intermediária, cujas funções de ativação são do tipo gaussiana (representadas pela equação (2.20)). Nelas, o parâmetro c define o centro da função gaussiana e σ^2 denota a sua variância, que indica a distância de um potencial de ativação u em relação ao centro c.

$$g(u) = e^{\left(\frac{-(u-c)^2}{2\sigma^2}\right)} \tag{2.20}$$

Portanto, os parâmetros a serem ajustados neste estágio são justamente o centro c e a variância σ_2 das funções gaussianas. No contexto da rede neural, "o centro c está diretamente associado aos seus próprios pesos, sendo que a entrada $u_j^{(1)}$ de cada um será o próprio vetor de entrada $\vec{x}$, que representa os n sinais externos que serão aplicados na rede" (SILVA; SPATTI; FLAUZINO, 2010). Consequentemente, a saída de cada neurônio j da camada intermediária é expressa pela equação (2.21).

$$g_j^{(1)}(u_j^{(1)}) = g_j^{(1)}(\mathbf{x}) = e^{\left(-\frac{\sum_{i=1}^{n}\left(x_i - w_{ji}^{(1)}\right)^2}{2\sigma_j^2}\right)}, \text{ onde } j = 1, \ldots, n_1 \tag{2.21}$$

Segundo Powell (1987), "quanto mais próxima uma determinada amostra estiver do centro de uma gaussiana, mais significante será o valor produzido pelo campo receptivo radial da função de ativação". Portanto, deve-se posicionar os centros das gaussianas da forma mais apropriada possível; isto pode ser feito com técnicas de agrupamento (*clustering*). Um destes algoritmos, chamado *k-means* (MACQUEEN, 1967), visa posicionar os centros de k gaussianas em regiões onde os padrões de entrada tenderão a se agrupar. No contexto da rede neural, "o valor do parâmetro k é igual ao número de neurônios da camada intermediária, já que a função de ativação de cada um é gaussiana" (SILVA; SPATTI; FLAUZINO, 2010).

Dado um conjunto de k gaussianas com centros em $\vec{W}_{1,j}^{(1)}$, $\vec{W}_{2,j}^{(1)}$, ..., $\vec{W}_{k,j}^{(1)}$, o algoritmo atribui cada amostra de treinamento x ao *cluster* $\Omega^{(i)}$ mais próximo de acordo com a medida da distância Euclidiana. Em seguida, os centroides dos *clusters* são movidos para posição média das amostras que foram atraídas a ele. Este processo é repetido até que não haja mais movimentações a serem feitas.

Finalmente, os campos receptivos produzidos pelos pesos dos neurônios associados aos *clusters* $\vec{\Omega}$ poderão ser obtidos calculando-se a variância das funções de ativação por meio da distância quadrática média, dada pela equação (2.22).

$$\sigma_j^2 = \frac{1}{|\Omega^{(j)}|} \sum_{x^{(k)} \in \Omega^{(j)}} \sum_{i=1}^{n} \left(x_i^{(k)} - W_{ji}^{(1)} \right)^2 \tag{2.22}$$

A figura 14 ilustra as iterações inicial, final e os raios de abrangência dos campos receptivos formados ao redor dos vetores de pesos ajustados pelo algoritmo *k-means*.

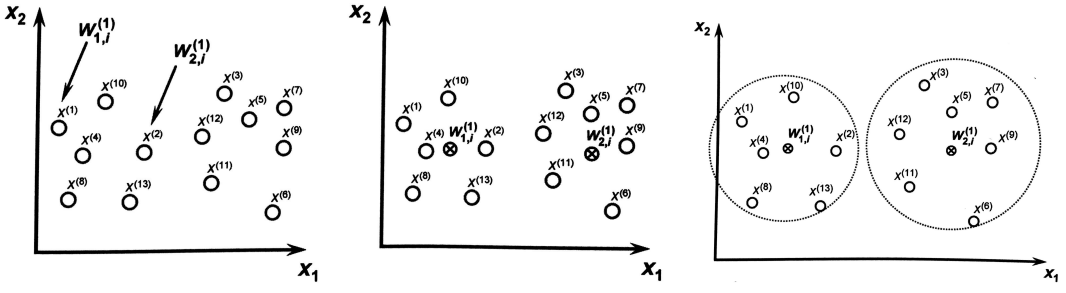


Figura 14: Iterações inicial, final e campos receptivos de uma rede RBF.

Etapa 2: treinamento supervisionado Como já foi dito, este estágio consiste em ajustar os pesos dos neurônios da camada de saída. Neste caso, diferentemente do estágio anterior, "o conjunto de treinamento será constituído por amostras rotuladas, em que as entradas serão as respostas das funções de ativação gaussianas dos neurônios da camada intermediária frente às respectivas amostras de treinamento" (SILVA; SPATTI; FLAUZINO, 2010), ou seja:

$$u_j^{(2)} = \sum_{i=1}^{n_1} W_{ji}^{(2)} \cdot g_i^{(1)}(u_i^{(1)}) - \theta_j, \text{ sendo } j = 1, \ldots, n_2 \text{ e:} \tag{2.23}$$

- $W_{ji}^{(2)}$ os pesos dos neurônios da camada de saída;
- θ_j os limiares dos neurônios da camada de saída;
- $g_i^{(1)}(u_i^{(1)})$ as saídas da camada intermediária, calculadas usando os valores de $W_{ji}^{(1)}$ e $\sigma_j^{(2)}$ obtidos no primeiro estágio de treinamento.

Basicamente, os valores dos pesos são ajustados em direção oposta ao gradiente, a fim de minimizar o erro, como pode ser visto na equação (2.24):

$$\Delta W_{ji}^{(2)} = \eta \cdot \frac{\partial E}{\partial W_{ji}^{(2)}} = \eta \cdot \left(d_j - g_j^{(2)}\right) \cdot g'\left(u_j^{(2)}\right) = \eta \cdot \left(d_j - g\left(u_j^{(2)}\right)\right) \cdot g'\left(u_j^{(2)}\right) \tag{2.24}$$

Na qual:

- $\Delta W_{ji}^{(2)}$ é o ajuste que será realizado nos pesos sinápticos da camada de saída;
- η é a taxa de aprendizado da rede, cujo valor situa-se entre 0 e 1;
- d_j é o valor esperado da j-ésima característica da amostra;
- $g_j^{(2)}$ é a saída do j-ésimo neurônio da camada de saída;
- $u_j^{(2)}$ é a entrada ponderada em relação ao j-ésimo neurônio de saída.

Finalmente, adotando-se uma função de ativação linear na camada de saída, seus neurônios realizarão uma combinação linear das gaussianas produzidas pelos neurônios da camada anterior. Portanto, a partir da equação (2.23) e com auxílio da figura 13, as respostas produzidas pelos neurônios de saída serão dadas por:

$$y_j = g_j^{(2)}(u_j^{(2)}) = u_j^{(2)}, \text{ sendo } j = 1, \ldots, n_2 \tag{2.25}$$

Desta forma, os estágios de treinamento ajustam os parâmetros da rede RBF.

2.6.3 Árvores de Decisão

Há diversas maneiras de representar classificadores e talvez a árvore de decisão seja a mais comum. Originalmente, fora estudada nos campos de teoria de decisão e estatística, mas sua efetividade em outras disciplinas, tais como aprendizado de máquina e reconhecimento de padrões, vem se tornando cada vez mais notável.

Árvores de decisão são também implementadas em várias aplicações do mundo real, e em Quinlan (1986) são mencionadas as três a seguir:

- Diagnóstico de condição médica a partir de um ou mais sintomas;
- Determinação do valor (na teoria dos jogos) de uma posição de xadrez;
- Opinar, a partir de observações atmosféricas, sobre a possibilidade de chover.

Uma árvore de decisão é um classificador expresso como uma partição recursiva do espaço de instâncias. Consiste em nós que formam uma árvore enraizada (ou seja, há um nó "raiz" que não possui arestas de entrada). Um nó com arestas de saída é chamado de interno ou nó de teste. Todos os outros nós são chamados de folhas (também conhecidos como nós terminais ou de decisão).

Em árvores de decisão, os nós com arestas de saída (chamados de internos ou de teste) dividem o espaço de instâncias em dois ou mais subespaços com base nos valores dos atributos de entrada (ROKACH; MAIMON, 2005):

- No caso de atributos discretos/categóricos, cada teste considera um único atributo, particionando o espaço de acordo com o valor do atributo.
- No caso de atributos contínuos/numéricos, a condição refere-se a um intervalo.

Os nós do tipo folha (também chamados de terminais ou de decisão) representam as possíveis saídas que uma instância pode assumir. A classificação de uma instância dá-se pela travessia da árvore, começando pela raiz e terminando em uma folha, de acordo com a saída dos testes ao longo do caminho.

A figura 15 ilustra uma árvore de decisão que, baseada em atributos de salário, tempo de deslocamento e oferta de café grátis, decide se uma proposta de emprego

será aceita. Nela, nós de teste são representados por losangos rotulados com o atributo que testa, folhas por elipses e as saídas como rótulos das conexões.

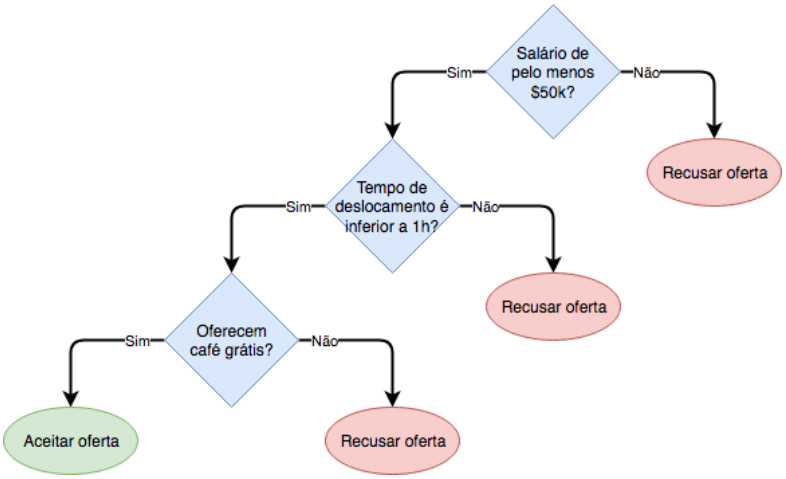


Figura 15: Árvore de decisão apresentando resposta à oferta de emprego.

Dispondo de um classificador deste tipo, um recrutador é capaz de prever a resposta de um candidato em particular e entender as características comportamentais de uma população inteira de candidatos.

O processo de criação de árvores de decisão a partir de um conjunto de dados de treinamento é chamado de indução. Diversas técnicas com esta finalidade foram propostas, tais como os algoritmos CART (BREIMAN et al., 1984), ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e ADTrees (FREUND; MASON, 1999).

Um aspecto importante sobre árvores de decisão é sua complexidade estrutural. De acordo com (BREIMAN et al., 1984), "tal característica (normalmente medida em termos do número total de nós, folhas, quantidade de atributos usados e profundidade total) tem um efeito crucial na precisão de classificação da árvore".

Há diversas técnicas que visam simplificar árvores excessivamente complexas. Os chamados algoritmos de poda visam aumentar tanto a compreensibilidade de uma árvore de decisão para um usuário humano quanto sua capacidade de generalização. Em Mingers (1989b), observou-se que, para diversos conjuntos de dados, a poda aumentou a precisão de classificação de 20% a 25%.

2.6.3.1 Reduced-Error Pruning Tree (REPT)

Como já foi apresentado na seção 2.6.3, a complexidade de uma árvore de decisão exerce forte influência no seu desempenho de classificação, e a forma mais comum de simplificação (visando aumentar a capacidade de generalização) é a poda, ou seja, remoção de subárvores que tornam sua estrutura desnecessariamente complexa (vide exemplo na figura 16).

(a) Árvore antes da poda. (b) Árvore depois da poda.

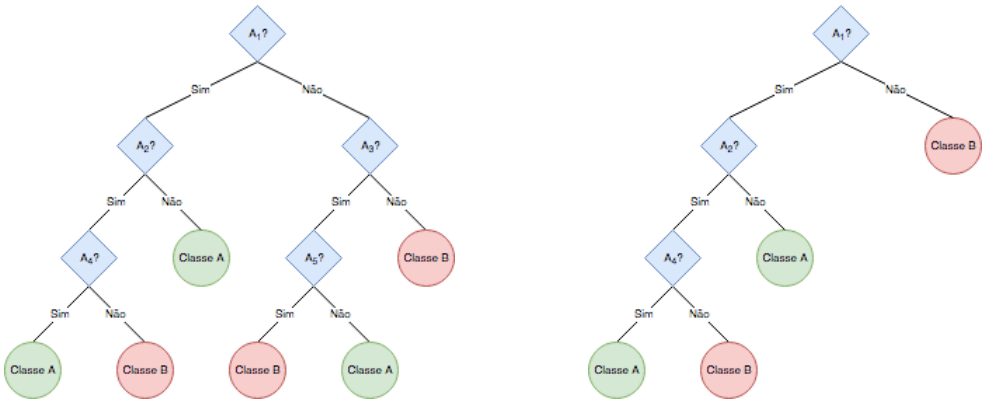


Figura 16: Exemplo de poda de árvore de decisão.

Diversos métodos de poda foram propostos na literatura. Nos trabalhos de Esposito, Malerba e Semeraro (1993) e Mingers (1989a), há revisões e estudos comparativos de vários deles, incluindo *Error-Complexity Pruning*, *Minimum-Error Pruning*, *Reduced-Error Pruning* e *Pessimistic Error Pruning*.

Neste trabalho, foi adotado o método *Reduced-Error Pruning* (QUINLAN, 1987), devido à sua simplicidade conceitual, velocidade de execução e eficácia na simplificação das árvores. Este tipo de poda funciona da seguinte forma:

```
V := conjunto de validação
Para cada nó N na árvore, faça:
  P(V) := precisão de classificação do conjunto de validação antes da poda em N
  Podar N, transformando-o em uma folha da classe mais comum em N
  P(V)' := precisão de classificação do conjunto de validação após a poda em N
  Se P(V)' < P(V), então desfazer a remoção da subárvore
  Caso contrário, então a poda de N foi realizada com sucesso
```

2.6.3.2 Boosted Reduced-Error Pruning Tree (AB-M1)

Boosting (em português "impulsionamento") é um método para melhorar o desempenho de qualquer algoritmo de aprendizado. Em teoria, ele pode ser usado para reduzir significativamente o erro de qualquer algoritmo de aprendizado "fraco", ou seja, que gere consistentemente classificadores que sejam apenas um pouco melhores do que adivinhações aleatórias. (SCHAPIRE, 1990) (FREUND, 1995)

Em Freund e Schapire (1996), são propostas duas versões do algoritmo: AdaBoost.M1, para problemas de classificação binária, e AdaBoost.M2, para problemas com mais de duas classes. Como este trabalho aborda justamente uma classificação binária, ou seja, atribuir um rótulo *ham* ou *spam* a um *e-mail*, o foco das explicações a seguir será na primeira versão do algoritmo.

O método, que está ilustrado na figura 17, funciona executando repetidamente um algoritmo de aprendizado "fraco" em diferentes distribuições sobre os dados de treinamento e, em seguida, combinando os classificadores $h_1, \ldots, h_n$ produzidos pelo algoritmo com seus respectivos pesos $\alpha_1, \ldots, \alpha_n$ formando assim um único classificador composto H.

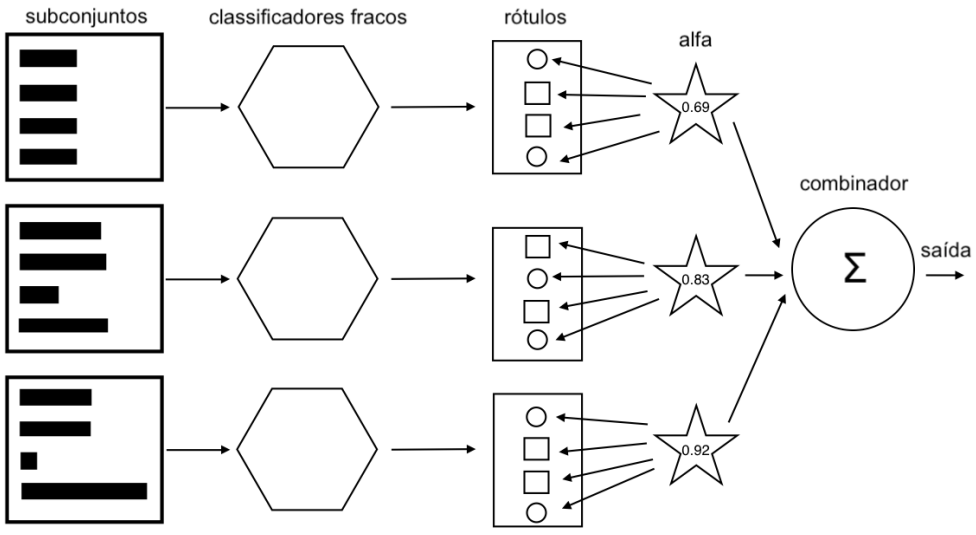


Figura 17: Ilustração do processo de funcionamento do AdaBoost.

Inicialmente, treina-se o primeiro classificador (h_1) atribuindo probabilidades iguais aos exemplos de um subconjunto aleatório de treinamento. Em seguida, é computado o peso α_1 deste classificador usando $t = 1$ na equação (2.26), na qual ϵ_t representa sua taxa de erro sobre o subconjunto.

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right) \tag{2.26}$$

Analisando a equação (2.26), notam-se os seguintes comportamentos de α para alguns valores de ϵ (MCCORMICK, 2013):

- α cresce positiva e exponencialmente à medida que ϵ se aproxima de 0.
- α é zero se ϵ for 0,5, pois um classificador com precisão de 50% não é melhor do que adivinhação aleatória e, portanto, deve ser ignorado.
- α cresce negativa e exponencialmente quando ϵ se aproxima de 1.

Após calcular α_1, os pesos dos exemplos de treinamento (representados pelo vetor $\vec{D}$) são atualizados usando a equação (2.27), na qual i é o índice de cada amostra.

$$D_{t+1}(i) = \frac{D_t(i) e^{(-\alpha_t y_i h_t(x_i))}}{\sum_i D_t(i)} \tag{2.27}$$

Cada valor D_i representa a probabilidade de o i-ésimo exemplo ser incluído no conjunto de treinamento. Entretanto, para que $\vec{D}$ represente uma distribuição, é necessário que $\sum_i D_i = 1$. Isto é feito via normalização, ou seja, dividindo cada D_i pelo somatório de todos os pesos, como mostra o denominador da equação (2.27).

Depois de treinar todos os T classificadores "fracos", a saída do classificador "forte" final é dada pela equação (2.28):

$$H(x) = \text{sgn} \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) \tag{2.28}$$

Neste trabalho, foi adotado como classificador "fraco" o *Reduced-Error Pruning Tree* (REPT) já explicado na seção 2.6.3.1, o que tornou possível avaliar se o AdaBoost.M1 foi capaz de prover uma melhora na precisão de classificação.

2.6.4 Máquina de Vetores de Suporte

Originalmente propostas por Vapnik e Lerner (1963), posteriormente aprimoradas por Boser, Guyon e Vapnik (1992) e sendo a formulação atualmente tomada como padrão a estabelecida por Cortes e Vapnik (1995), as máquinas de vetores de suporte consistem em conjuntos de modelos supervisionados de aprendizado e algoritmos associados que analisam dados usados para classificação e regressão.

No contexto da classificação, dado um conjunto de exemplos de treinamento rotulados, um algoritmo de treinamento de SVM constrói um modelo capaz de atribuir classes a novos exemplos, com a propriedade de separar categorias distintas por uma lacuna o mais larga possível (vide figura 18), no intuito de minimizar o erro de generalização do classificador.

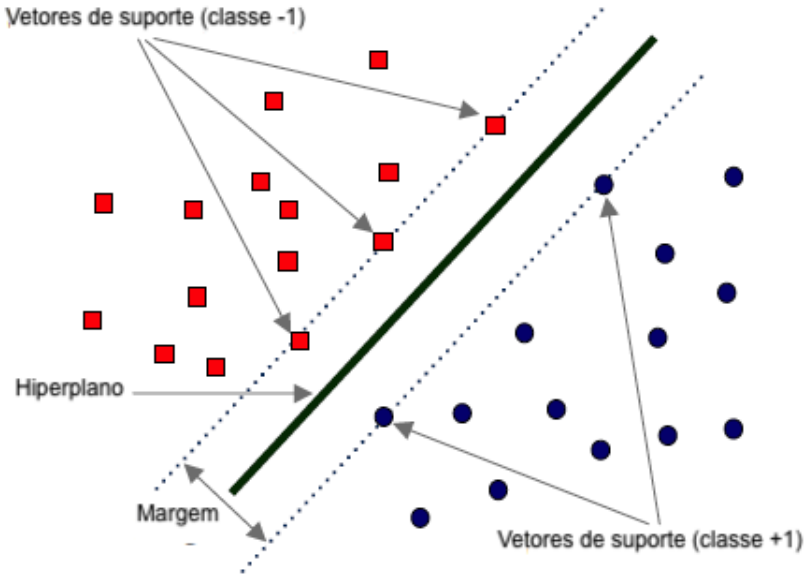


Figura 18: Hiperplano, vetores de suporte e margem de separação.

Na figura 18 é possível notar que há amostras de classes distintas em cada extremo da margem de separação. Cada amostra $(\vec{x}, y)$ é uma tupla composta por um ponto n-dimensional $\vec{x}$ e uma saída associada y. Justamente por "apoiarem" a margem de separação, estes dados situados nas fronteiras são chamados de vetores de suporte, o que dá nome também à técnica SVM.

Além de classificação linear, as SVMs podem realizar classificação não-linear de forma eficiente usando o chamado *kernel trick*. Esta técnica mapeia o espaço original de entradas (em que as classes não são linearmente separáveis) em espaços de características de igual ou mais alta dimensionalidade (vide figura 19), nos quais a separabilidade linear é possível. Para efetuar tal transformação, é usada uma função de *kernel*.

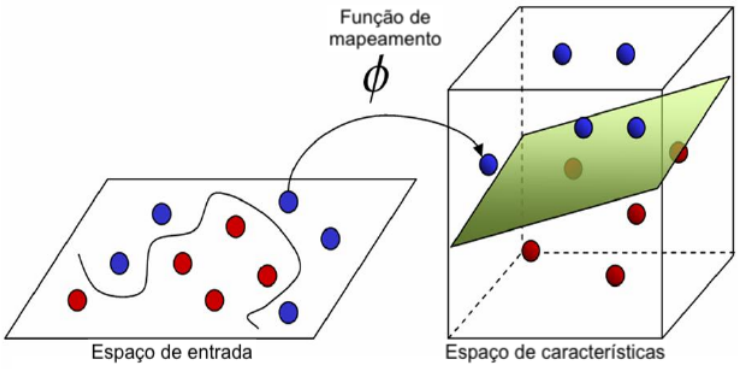
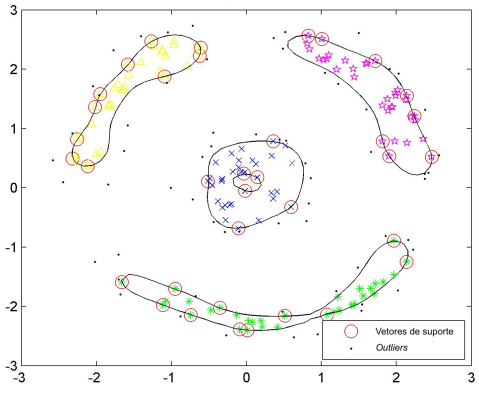


Figura 19: Aumento de dimensionalidade por meio do *kernel trick*.

As SVMs também conseguem encontrar agrupamentos de dados quando as amostras de treinamento não são rotuladas. Para isto, podem ser adotadas técnicas baseadas em grafos completos (BEN-HUR et al., 2001) ou de proximidade (YANG; ESTIVILL-CASTRO; CHALUP, 2002) (vide figuras 20a e 20b, respectivamente).

(a) Agrupamento com grafo completo.



(b) Agrupamento com grafo de proximidade.

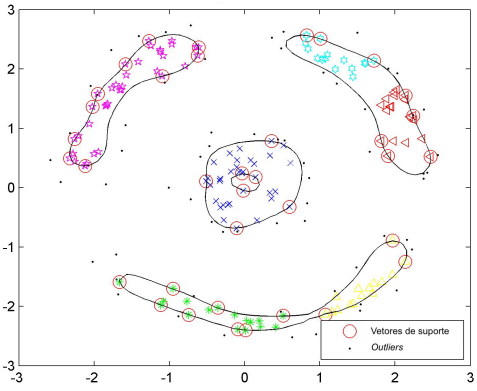


Figura 20: Agrupamentos de dados usando máquinas de vetores de suporte.

2.6.4.1 Linear Support Vector Machine (L-SVM)

Como dito na seção 2.6.4, o conjunto de treinamento utilizado pelas SVMs em cenários de classificação são tuplas $(\vec{x}, y)$ compostas por um ponto n-dimensional $\vec{x}$ e uma saída associada y. Neste sentido, a construção do modelo consiste em encontrar um hiperplano capaz de separar as amostras de treinamento de forma que haja a maior distância possível entre as categorias do problema.

Quando os dados de treinamento são linearmente separáveis, é possível encontrar um hiperplano de "margem rígida" que atenda ao requisito de separabilidade máxima desde que não haja erros e/ou *outliers* nos dados. Entretanto, caso os dados não sejam linearmente separáveis e/ou deseja-se permitir certa tolerância de erros de classificação (visando evitar *overfitting*), deve-se então encontrar um hiperplano de "margem suave", que não se ajustará totalmente e apresentará capacidade de generalização potencialmente superior.

(a) SVM de margem rígida. (b) SVM de margem suave.

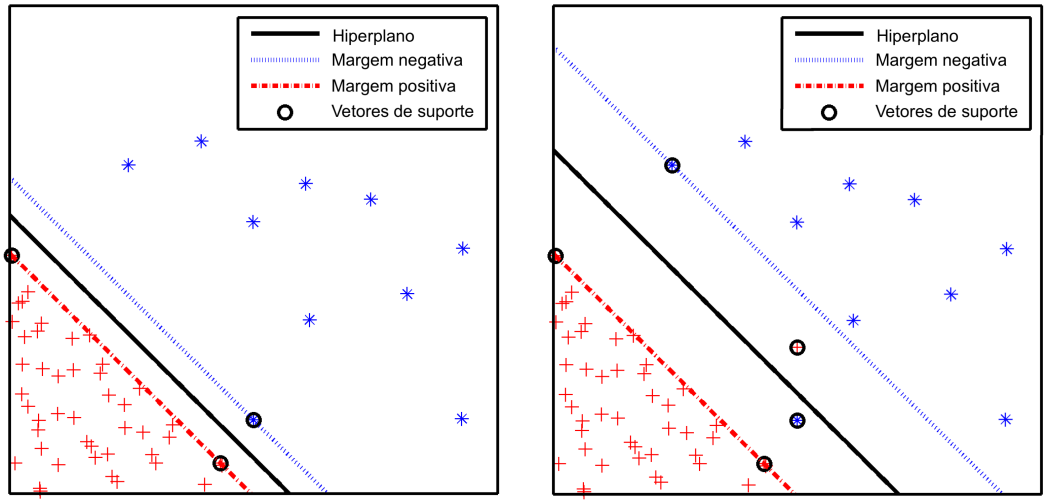


Figura 21: Representação gráfica das margens de separação rígida e suave.

Nas figuras 21a e 21b estão ilustradas SVMs de margem rígida e suave, respectivamente. Na primeira não há pontos com classificação incorreta; entretanto, a margem de separabilidade é estreita. Já na segunda, há uma amostra de cada classe classificada incorretamente; porém, a margem de separabilidade é larga.

A “margem” é a região delimitada por um par de hiperplanos paralelos que separam as duas classes de dados com a maior distância possível, e o “hiperplano de margem máxima” é aquele que fica no meio do caminho entre eles.

Para encontrar tal hiperplano em casos linearmente separáveis, deve-se resolver o seguinte problema de otimização: “dado o conjunto de treinamento $\{(\vec{x}_i, d_i)\}_{i=1}^{N}$, encontre os valores ótimos do vetor peso $\vec{w}$ e *bias* b de modo que satisfaçam as restrições $d_i \left(\vec{w}^T \vec{x}_i + b \geq 1\right)$ para $i = 1, 2, \ldots, N$ e o vetor peso $\vec{w}$ minimize a função de custo $\Phi(\vec{w}) = \frac{1}{2}\vec{w}^T\vec{w}$” (HAYKIN, 2001).

Em casos não linearmente separáveis, são introduzidas as chamadas variáveis soltas ξ_i, que medem o desvio de um ponto de dado da condição ideal de separabilidade de padrões. Desta forma, obtém-se o seguinte problema de otimização: “dado o conjunto de treinamento $\{(\vec{x}_i, d_i)\}_{i=1}^{N}$, encontre os valores ótimos do vetor peso $\vec{w}$ e *bias* b de modo que satisfaçam às restrições $d_i \left(\vec{w}^T \vec{x}_i + b \geq 1 - \xi_i\right)$ para $i = 1, 2, \ldots, N$ e $\xi_i \geq 0$ para todo i e de modo que o vetor peso $\vec{w}$ e as variáveis soltas ξ_i minimizem a função de custo $\Phi(\vec{w}, \xi) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{N}\xi_i$ em que C é um parâmetro positivo especificado pelo usuário” (HAYKIN, 2001).

O parâmetro C indica o quanto deseja-se evitar erros de classificação. Portanto, valores grandes implicam na escolha de um hiperplano de margem menor, enquanto valores pequenos resultarão em hiperplanos com margem maior. A figura 22 ilustra a influência deste parâmetro na fronteira de decisão de uma SVM.

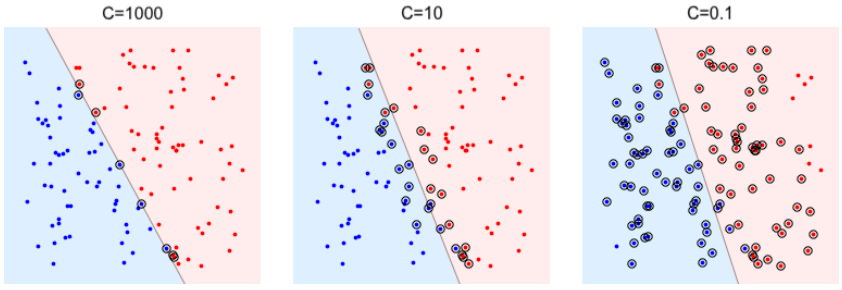


Figura 22: Ilustração da influência do parâmetro C na margem de uma SVM.

Tanto no caso linear quanto no não-linear, os $\vec{w}$ e b que solucionam os problemas de otimização determinam o classificador $\vec{x} \mapsto \text{sgn}(\vec{w} \cdot \vec{x} - b)$.

2.6.4.2 Non-Linear Support Vector Machine (NL-SVM)

Na seção 2.6.4.1, foram apontados os problemas de otimização cujas soluções resultam em SVMs de margem rígida e suave. Como dito, o primeiro tipo destina-se a conjuntos de dados linearmente separáveis (sobre os quais não são tolerados erros de classificação), enquanto o segundo é normalmente utilizado em conjuntos de dados não-linearmente separáveis (sobre os quais podem haver erro de classificação de modo a preservar a capacidade de generalização do classificador).

Em ambos os casos, é importante ressaltar que o classificador produzido é de natureza linear, por trabalhar com hiperplanos construídos no espaço original de características e por utilizar como função de *kernel* o produto interno. A ideia de criar classificadores não-lineares foi proposta em Boser, Guyon e Vapnik (1992), cuja abordagem consiste na aplicação da *kernel trick* (AIZERMAN; BRAVERMAN; ROZONOÉR, 1964) aos hiperplanos de margem máxima.

Uma SVM não-linear funciona de forma similar a uma linear; entretanto, os produtos internos são substituídos por uma função de *kernel* não-linear (vide alguns exemplos na tabela 4 (HAYKIN, 2001)). Isto permite ao algoritmo ajustar o hiperplano de margem máxima em um espaço de características transformado de igual ou maior dimensionalidade e linearmente separável. Somente a máquina de vetores de suporte de *kernel* polinomial (P-SVM) foi usada nos experimentos.

Tabela 4: Exemplos de funções de *kernel*.

Tipo	Função
Linear	$K(\vec{x}, \vec{x_i'}) = \vec{x}^T \vec{x}_i$
Polinomial	$K(\vec{x}, \vec{x_i'}) = (\vec{x}^T \vec{x}_i)^d$
Gaussiana	$K(\vec{x}, \vec{x_i'}) = \exp\left(-\frac{1}{2\sigma^2}\|\|\vec{x} - \vec{x_i}\|\|^2\right)$
Sigmoide	$K(\vec{x}, \vec{x_i'}) = \tanh(\eta \vec{x} \vec{x_i} + \theta)$

Outras abordagens que utilizam a *kernel trick* incluem a categorização de textos com base em subsequências de caracteres (LODHI et al., 2002) e o uso de *kernels* compostos na classificação de imagens espectrais (CAMPS-VALLS et al., 2006).

3 Revisão Bibliográfica

O estudo e desenvolvimento de sistemas anti-*spam* são assuntos recorrentes em trabalhos da área de Ciência da Computação. Pesquisadores vêm abordando o tema das mais diversas formas, propondo soluções que utilizam desde técnicas estatísticas e redes neurais artificiais à análise da reputação dos remetentes e detecção de padrões e estilos de escrita frequentes em mensagens indesejadas.

O objetivo do presente trabalho é associar técnicas de seleção de características, redução de dimensionalidade e modelos de aprendizado de máquina de modo a construir um classificador de *e-mails* automático, robusto e adaptável. Portanto, neste capítulo são apresentados e discutidos (em ordem cronológica de data de publicação) outros trabalhos e pesquisas envolvendo classificação de *e-mails*.

Drucker et al. (1999) comparam as SVMs a três outras técnicas (Boosting, Ripper e Rocchio) aplicadas à classificação de *e-mails*. Os testes foram conduzidos em dois conjuntos de dados oriundos da companhia americana de telecomunicações AT&T, submetidos a múltiplos métodos de seleção de características (binário ou baseados em frequências dos termos) e a uma filtragem de palavras vazias. Além de resultados de precisão em torno de 98%, os autores destacaram características importantes da SVM: a melhora de velocidade de treinamento ao usar características binárias[1], a sensibilidade ao parâmetro C[2] e a capacidade de lidar com grandes

[1]Neste caso, a maioria das características é nula, resultando em vetores esparsos cujo produto interno pode ser otimizado com rotinas não-multiplicativas.

[2]Parâmetro de regularização que controla o balanço entre alcançar um erro baixo nos dados de treinamento e minimizar a norma dos pesos.

quantidades de características (em vez de subconjuntos).

Utilizando o classificador probabilístico *Naïve Bayes* aliado a técnicas linguísticas de lematização[3] e remoção de palavras vazias[4], Androutsopoulos et al. (2000a) obtiveram taxas de precisão de classificação superiores a 95% no *corpus* Ling Spam (ANDROUTSOPOULOS et al., 2000b), com um custo computacional relativamente baixo. No trabalho, o autor observa que, apesar de a adoção das técnicas linguísticas ter melhorado sensivelmente a precisão na classificação, elas introduzem novas possibilidades na análise das mensagens. Por fim, foram realizados experimentos com diferentes custos atribuídos a falsos-positivos e falsos-negativos. Este aspecto experimental levou à conclusão de que este modelo é inviável nas situações em que *e-mails* são descartados quando identificados como *spam*.

Meyer e Whateley (2004) avaliaram o desempenho de um classificador *Naïve Bayes* usado em conjunto com esquemas de combinação (probabilidades χ^2 a fim de pontuar mensagens como *ham*, *spam* ou incertas), n-gramas[5] (uni ou bi-gramas) e *tokenização* (símbolos de cabeçalhos, corpo ou "sintéticos" — características indiretas dos *e-mails*, como ausência de remetente, assunto etc.). O treinamento foi feito sobre cinco *corpora* — quatro SpamBayes e o Spam Assassin (The Apache Software Foundation, 2005), utilizando um regime de treinamento configurável (que indica ao algoritmo "quando" a aprendizagem deve ser feita: com todas as mensagens, somente com mensagens dentro de uma faixa de pontuação etc.) e validação cruzada. A quantidade de mensagens com necessidade de classificação manual após o treinamento foi de apenas pouco mais de 1%, comprovando a eficácia da combinação de técnicas escolhida.

Berger e Merkl (2005) realizaram um estudo comparativo das técnicas de aprendizado supervisionado SVM e não-supervisionado SOM (do inglês *Self-Organizing Map*) aplicadas à categorização de *e-mail*. Nos experimentos, as características

[3]Representação de palavras em sua forma básica, desconsiderando mudanças gramaticais como o tempo e a pluralidade.

[4]Palavras que são filtradas antes ou depois do processamento de dados de linguagem natural (texto); geralmente referem-se às palavras mais comuns em um idioma.

[5]Subsequência de n elementos de uma dada sequência.

dos *e-mails* (oriundos de um conjunto pessoal de mensagens) foram representadas tanto como palavras quanto como n-gramas, sob a justificativa destes apresentarem maior robustez em relação a erros de ortografia e independerem de contexto. Nos resultados experimentais, ambas as técnicas apresentaram precisão superior a 90%, tendo a SOM resultados levemente inferiores à SVM, superando as expectativas dos pesquisadores devido à natureza não-supervisionada de seu treinamento.

Carpinteiro et al. (2006) propuseram o pré-processamento de *e-mails* visando simplificar as mensagens e melhorar o levantamento de suas características. Analisando diversas partes dos *e-mails* (cabeçalho, corpo — tanto texto quanto HTML[6] — e tipos dos anexos) e aplicando diferentes métodos de seleção de características, foi treinada uma rede neural do tipo MLP (do inglês *Multilayer Perceptron*). Os experimentos mostraram precisão de classificação próxima a 100% na base Spam Assassin (The Apache Software Foundation, 2005). Para trabalhos futuros, são sugeridos outros modelos neurais, formas de pré-processamento de mensagens e métodos de seleção de características.

Sirisanyalak e Sornil (2007) propuseram uma solução para classificação de *e-mails* baseada em um sistema imunológico artificial. Simulando a resposta imunológica biológica, o sistema deve ser capaz tanto de gerar e clonar anticorpos[7] eficazes quanto de detectar e eliminar anticorpos ineficazes (ou seja, aqueles que atacam células do próprio organismo). Submetendo o *corpus* Spam Assassin (The Apache Software Foundation, 2005) a uma análise sintática de MIME[8], remoção de palavras vazias e lematização, um conjunto de detectores[9] é gerado, aprimorado e sintetizado por meio de algoritmos genéticos. Finalmente, é realizada a extração de quatro características específicas[10] e a base é submetida a avaliações previamente

[6]Linguagem de marcação utilizada na construção de páginas na *Web*.

[7]Unidades dedicadas ao reconhecimento de um ou mais antígenos (neste caso, características de *spams*) por meio de propriedades particulares a estes.

[8]Do inglês *Multipurpose Internet Mail Extensions*; padrão da internet para o formato das mensagens de correio eletrônico.

[9]Neste caso, consistem em expressões regulares capazes de detectar palavras e suas variações causadas por técnicas de ofuscação, por exemplo.

[10]Número de termos na mensagem, de vezes que a mensagem casa com o conjunto de detectores,

submetidas a métodos de escolha de parâmetros e realizadas por uma função de regressão logística[11], obtendo resultados de precisão superiores a 98%.

Zhang et al. (2009) propuseram um sistema de classificação de *e-mails* baseado em reputação, no qual um servidor armazena grupos de endereços IP (em uma estrutura de dados do tipo árvore) e suas respectivas pontuações. Compostas a partir do histórico de envio de *e-mails* destes grupos de usuários e realimentadas tanto por outros sistemas anti-*spam* quanto pelos destinatários, as notas indicam o quanto um remetente é confiável. Utilizando um *corpus* composto por quase 3 milhões de *e-mails* oriundos de um servidor universitário, os experimentos mostraram que o sistema proposto atingiu mais de 95% de acurácia, precisão e revocação.

Pérez-Díaz et al. (2012) propuseram o uso da técnica de conjuntos aproximados (PAWLAK, 1991) na categorização de *e-mails* aliada a heurísticas[12] para endereçar o problema de mensagens com classificação duvidosa. Após submeter a base Spam Assassin (The Apache Software Foundation, 2005) a uma remoção de palavras vazias e utilizando representações binárias ou em termos de frequência, os autores executam a técnica de conjuntos aproximados em regime de validação cruzada e comparam os resultados de classificação aos das técnicas AdaBoost, Flexible Bayes, Naïve Bayes e SVM, usando diferentes quantidades de características para cada. O método obteve taxas de escore F_1 de 98%, superando os demais modelos adotados na comparação. Por fim, os autores destacam a importância de regenerar periodicamente o conjunto de regras derivado pela técnica devido ao *concept drift*[13] e alertaram para o grande tempo de treinamento, sugerindo a adoção de métodos de redução do espaço de características.

---

de detectores que casam com a mensagem e um fator de tendência de *spam*.

[11] Técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias.

[12] MFD (do inglês *Most Frequent Decision*, LNO (do inglês *Largest Number of Objects*) e LTS (do inglês *Largest Total Strength*).

[13] Noção de que propriedades estatísticas da variável alvo, das quais o modelo está tentando prever, mudam ao longo do tempo de maneira imprevisível, tornando as previsões precisas à medida que o tempo passa.

No trabalho de Yevseyeva et al. (2013), três algoritmos evolutivos são aplicados à otimização de filtros anti-*spam* (sendo um deles de único objetivo — Grindstone4SPAM (MÉNDEZ et al., 2012) — e os outros dois, multiobjetivos — NSGA-II (do inglês *Nondominated Sorting Genetic Algorithm II*) (DEB et al., 2002) e SPEA2 (do inglês *Strength Pareto Evolutionary Algorithm 2*) (ZITZLER; LAUMANNS; THIELE, 2001)) e comparados em termos de métricas específicas aos domínios de algoritmos estocásticos — EAF (do inglês *empirical attainment function* (LÓPEZ-IBÁÑEZ; PAQUETE; STÜTZLE, 2010)) — e de pesquisa de sistemas anti-*spam*— precisão, revocação, escore F_1 etc. Foram elaboradas duas estratégias de avaliação para abordar estes domínios, respectivamente: traçado de gráficos dos *fronts* de Pareto[14] (a fim de facilitar a visualização da eficiência de cada método evolutivo) e o treinamento e avaliação da influência da otimização em um regime de validação cruzada. Por fim, os resultados apontaram que o NSGA-II obteve o melhor desempenho, e os autores destacam que a variedade de soluções obtidas pelos algoritmos multiobjetivos não foi totalmente explorada no trabalho justamente pelas restrições impostas pelo arcabouço necessário à comparação de algoritmos de tipos diferentes.

Barigou, Beldjilali e Atmani (2014) propuseram uma melhoria ao algoritmo KNN (do inglês *K-Nearest Neighbors*), introduzindo autômatos celulares para reduzir a quantidade de documentos selecionados ao classificar uma amostra. Tal alteração conferiu uma redução de uso de memória[15] e aumento do desempenho[16] do classificador que, quando exercitado com a base Ling Spam (ANDROUTSOPOULOS et al., 2000b), apresentou precisão superior a 98%. Para trabalhos futuros, os autores sugerem avaliar a mesma técnica com a base Spam Assassin (The Apache Software Foundation, 2005) e estabelecer mais critérios de avaliação.

Youn (2014) propôs um filtro de *e-mails* de dois níveis baseado em ontologia[17]. A partir de um determinado conjunto de dados, as características são selecionadas

[14]Conjunto de soluções não-nomeadas, sendo escolhido como ótimo, se nenhum objetivo pode ser melhorado sem sacrificar pelo menos um outro objetivo.

[15]Devido à estrutura celular em que os dados do conjunto treinamento são organizados.

[16]Devido à redução do nível de ruído dos subconjuntos e à resistência a desbalanceamento.

[17]Representações formais de um conjunto de conceitos dentro de um domínio e suas relações.

pelo método TF-IDF (do inglês *Term Frequency–Inverse Document Frequency*)[18] e uma árvore de decisão C4.5 (QUINLAN, 1993) é induzida para posteriormente ser convertida para o formato RDF (do inglês *Resource Description Framework*) (CYGANIAK; WOOD; LANTHALER, 2014) e usada para produzir uma ontologia global (primeiro nível). As mensagens do conjunto de testes são objeto de consulta para esta ontologia, que fornece uma classificação baseada nos atributos das mensagens e em seu próprio modelo, que é periodicamente adaptado e pode ser usado para conceber uma ontologia personalizada (segundo nível). Submetido a um conjunto de dados particular, o método apresentou precisões em torno de 91% e 95% usando os filtros global e personalizado, respectivamente.

Visando melhorar o desempenho do algoritmo de seleção negativa, Idris et al. (2015) propuseram um modelo híbrido denominado NSA-PSO (do inglês *Negative Selection Algorithm — Particle Swarm Optimization*), substituindo o gerador de detectores aleatórios pelo algoritmo de otimização por enxame de partículas (J; C, 1995), usando como função de *fitness* a LOF (do inglês *Local Outlier Factor*) (BREUNIG et al., 2000). Os autores também propuseram um modelo de arquitetura modular para o sistema anti-*spam*, facilitando o caminho para uma implantação real. Por fim, experimentando diferentes configurações de limiar e quantidade de detectores, o modelo foi exercitado com o *corpus* Spambase (HOPKINS et al., 1999) e apresentou taxa de precisão de 77% e escore F_1 de 83% (contra respectivos 69% e 36% do algoritmo original), comprovando uma melhora de desempenho.

Kaya e Ertuğrul (2016) propuseram um novo método de seleção de características baseado na "probabilidade do uso de caracteres de ordens similares com respeito aos seus valores UTF-8 (do inglês *8-bit Unicode Transformation Format*) [19]", determinados pelo operador de baixo nível *shifted 1D-LBP* (do inglês *One-*

[18] Medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos.

[19] Tipo de codificação binária de comprimento variável capaz de representar qualquer caractere universal padrão do *Unicode* e ASCII (do inglês *American Standard Code for Information Interchange*).

Dimension Linear Binary Pattern)[20]. Nos experimentos, o método é aplicado aos conjuntos Ling Spam (ANDROUTSOPOULOS et al., 2000b), Spam Assassin (The Apache Software Foundation, 2005) e TREC (CORMACK; LYNAM, 2006) e, em seguida, alguns modelos de aprendizado de máquina[21] são treinados com validação cruzada e avaliados em termos de diversas métricas[22]. Os resultados de acurácia de classificação se mostraram promissores, atingindo aproximadamente 92%, 93% e 95% para os conjuntos de dados mencionados anteriormente, respectivamente.

Shams e Mercer (2016) propuseram a extração e o uso de atributos de estilometria[23] para treinar classificadores de *e-mails*. Como exemplos destes atributos, pode-se citar a quantidade de erros ortográficos e gramaticais, indicadores de facilidade de leitura (índice Gunning fog), SMOG (do inglês *Simple Measure of Gobbledygook*), FRES (do inglês *Flesch Reading Ease Score*), Forcast e legibilidade de Flesch-Kincaid), quantidades de palavras simples (com até duas sílabas) e complexas (com três ou mais sílabas) e tamanho médio de *e-mail* e de palavras. Aliados a atributos clássicos como IG (do inglês *Information Gain*) e TF-IDF, o método foi testado com os classificadores NB, RF e SVM e os *ensembles*[24] Bagging e Adaboost.M1 sobre os *corpora* CSDMC2010 (TAO, 2010), Spam Assassin (The Apache Software Foundation, 2005), Ling Spam (ANDROUTSOPOULOS et al., 2000b) e Enron-Spam (COHEN, 2008). Os métodos Bagging e Adaboost.M1 obtiveram os melhores resultados, e concluiu-se que a técnica é relevante na detecção de *spam* em *corpora* personalizados (ou seja, naqueles em que a coleta de *e-mails* não é aleatória), mas limitada em *corpora* não-personalizados, devido à ausência de padrões únicos de escrita neste tipo de conjunto de dados.

[20]Para uma dada mensagem, o operador é definido como um conjunto de comparações binárias entre o valor central e os valores vizinhos, resultando em um valor inteiro que representa um padrão de informação estrutural.

[21]FLDA (do inglês *Fisher Linear Discriminant Analysis*), NB, BN (do inglês *BayesNet*), FT (do inglês *Functional Tree*), RT (do inglês *Random Tree*) e RF (do inglês *Random Forest*).

[22]Precisão, revocação, acurácia, MCC (do inglês *Matthews Correlation Coefficient*), escore F_1 e ROC (do inglês *Receiver Operating Characteristic*).

[23]Aplicação do estudo do estilo linguístico, geralmente à linguagem escrita, frequentemente usado para atribuir autoria a documentos anônimos ou contestados.

[24]Uso de vários algoritmos de aprendizagem para obter melhor desempenho preditivo do que poderia ser obtido de qualquer um dos algoritmos de aprendizado constituintes sozinho.

Yang et al. (2016) propuseram um filtro anti-*spam* baseado no algoritmo IB (do inglês *Information Bottleneck*) destinado a conjuntos de treinamento pequenos, sob a justificativa de que a frequente mudança de temas dos *e-mails* reduz a disponibilidade imediata de amostras com contextos atuais. Basicamente, o algoritmo extrai amostras altamente significantes do conjunto, a fim de construir centroides usados para classificar mensagens em *ham* ou *spam* por meio de uma função de similaridade (neste caso, a divergência de Jensen–Shannon (LIN, 1991)). Submetendo os *corpora* Ling Spam (ANDROUTSOPOULOS et al., 2000b), Spambase (HOPKINS et al., 1999), PU3 (ANDROUTSOPOULOS; PALIOURAS; MICHELAKIS, 2003) e TREC (CORMACK; LYNAM, 2007) a uma extração de características baseadas em IG, o classificador apresentou resultados (de precisão, revocação e escore F_1) comparáveis aos competidores (SVM, NB e AdaBoost) quando trabalhando com conjuntos de treinamento grandes. Com a redução da quantidade de amostras, o algoritmo teve menor deterioração de desempenho, confirmando sua capacidade de predição mesmo com conjuntos pequenos.

Kumaresan, Saravanakumar e Balamurugan (2017) propuseram um *framework* para classificação de *e-mails* baseado em SVM. Combinando os *corpora* Ling Spam (ANDROUTSOPOULOS et al., 2000b) e Spam Archive (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007a) (DREDZE; GEVARYAHU; ELIAS-BACHRACH, 2007b), o conjunto de dados utilizado é composto por mensagens com texto e imagens, cujas características textuais são selecionadas pelo método TF (do inglês *Term Frequency*) e as visuais pelos métodos correlograma[25] e *wavelet moment*[26]. Após a extração, o espaço de características é reduzido por uma versão modificada do algoritmo de busca *Cuckoo search* com voos de Lévy[27] (YANG; DEB, 2010) e os vetores são submetidos à classificação realizada por uma SVM de *kernel* híbrido[28]. Após os experimentos, o método obteve taxa precisão de classificação superior a

[25]Gráfico utilizado em séries temporais para traçar autocorrelações, ou seja, a correlações entre uma série e ela mesma defasada.

[26]Técnica usada para medir a regularidade local de um sinal.

[27]Sucessão de passos aleatórios cujos comprimentos seguem uma distribuição de probabilidade de cauda pesada (exemplos: distribuições de Pareto, Lévy, Cauchy, Burr e t de Student).

[28]Utiliza uma combinação de dois ou mais *kernels*, como linear, polinomial e quadrático.

97%, superando os 94% obtidos pelos demais modelos adotados na comparação e evidenciando o aumento de desempenho proporcionado pela técnica proposta.

Os trabalhos revisados nesta seção propuseram diversas abordagens e modelos para categorização de texto e detecção de *spam*. Os modelos foram avaliados sobre diversos *corpora* e frequentemente obtiveram acerto de classificação superior a 90%. Entretanto, a redução de dimensionalidade do espaço de características das mensagens não foi explorada (no máximo, fora citada por (PÉREZ-DÍAZ et al., 2012) como sugestão para continuação de sua pesquisa) e, portanto, o modelo proposto difere-se dos demais neste aspecto.

4 Metodologia

4.1 Modelo proposto

A fim de avaliar o desempenho dos diferentes algoritmos de aprendizado aliados às técnicas de seleção de característica, redução de atributos e balanceamento de classes explicadas no capítulo 2, foi elaborado um modelo ilustrado no diagrama de blocos da figura 23, cujos elementos são explicados individualmente na sequência.

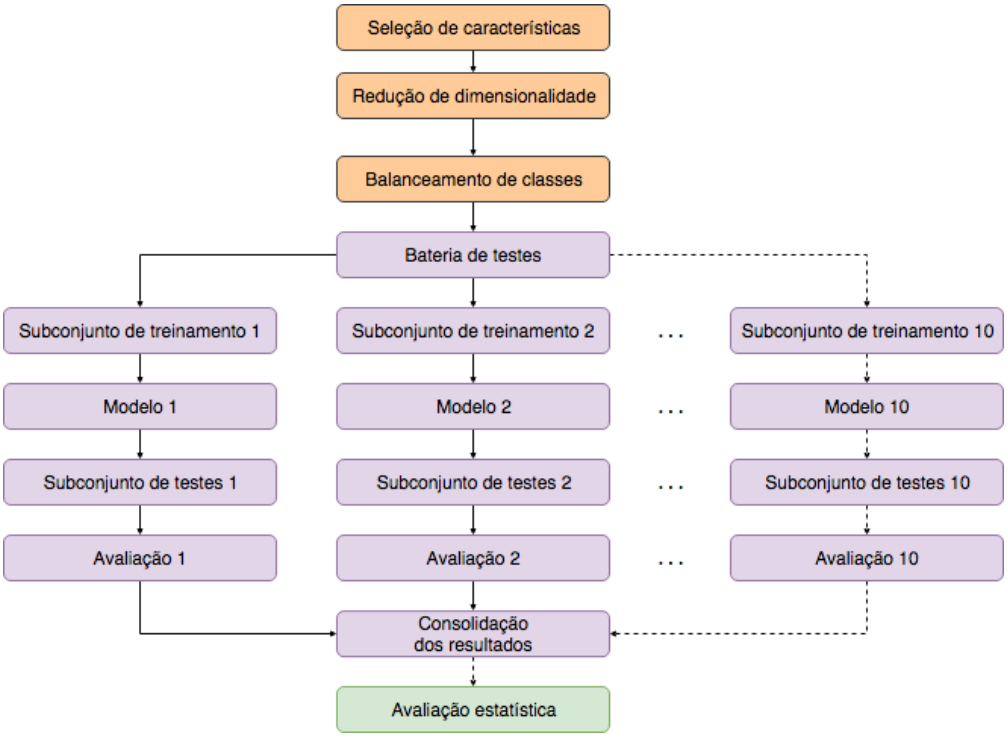


Figura 23: Diagrama de blocos da metodologia proposta.

- **Seleção de características**: Etapa que analisa os *e-mails* dos conjuntos de dados em sua forma textual e extrai suas características de acordo com os métodos descritos em 2.3, gerando uma representação numérica de cada mensagem, que será utilizada pelos modelos de aprendizado de máquina.
- **Redução de dimensionalidade**: Usando o método explicado na seção 2.4, é efetuada uma redução de dimensionalidade no conjunto de dados gerado na etapa anterior, preservando os atributos mais significativos de cada amostra e descartando os menos relevantes perante a classe de cada instância.
- **Balanceamento de classes**: Os conjuntos de dados já reduzidos são balanceados, a fim de equilibrar a contribuição de cada classe no treinamento do modelo. Esta técnica é descrita mais detalhadamente na seção 2.5.
- **Bateria de testes**: Nesta etapa, o conjunto de dados é embaralhado e divido ao meio[1], e as metades são destinadas às etapas de treinamento e classificação (aqui, os padrões zerados[2] mencionados na tabela 6 são incluídos no subconjunto). Em seguida, cada modelo é treinado e avaliado 10 vezes.
- **Subconjunto de treinamento**: Consiste na primeira metade do conjunto de dados reduzido e embaralhado.
- **Modelo**: É o artefato resultante do treinamento do modelo de aprendizado. Pode ser persistido no sistema de arquivos e recuperado para uso posterior.
- **Subconjunto de testes**: Consiste na segunda metade do conjunto de dados reduzido e embaralhado, somada aos padrões zerados.
- **Avaliação**: O modelo é avaliado por meio da classificação do subconjunto de testes, e o resultado de cada métrica (explicadas na seção 4.4) é calculado.
- **Consolidação**: As métricas calculadas na etapa anterior são consolidadas para que as médias e intervalos de confiança sejam calculados.
- **Avaliação estatística**: Os resultados dos 10 experimentos são consolidados para que a média e o intervalo de confiança sejam calculados e apresentados.

[1]Conjuntos de treinamento maiores que os de teste representam cenários irreais, pois a quantidade de dados disponível para pesquisa e aquisição de conhecimento é consideravelmente inferior àquela processada diariamente por grandes servidores de *e-mail* em ambientes reais.

[2]Mensagens cuja representação vetorial só possui valores nulos (ou seja, iguais a zero).

4.2 Ferramentas utilizadas

O projeto WEKA (do inglês *Waikato Environment for Knowledge Analysis*) (HALL et al., 2009) visa fornecer uma coleção abrangente de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados para pesquisadores e profissionais. Ele permite que os usuários experimentem rapidamente e comparem diferentes métodos de aprendizagem de máquina em novos conjuntos de dados. Sua arquitetura modular e extensível permite que processos sofisticados de mineração de dados sejam construídos a partir da ampla coleção de algoritmos e ferramentas de aprendizagem disponíveis. A extensão do *kit* de ferramentas é fácil graças a uma API (do inglês *Application Programming Interface*) simples, mecanismos de *plugin* e instalações que automatizam a integração de novos algoritmos de aprendizagem com as interfaces gráficas de usuário do WEKA (vide figura 24).

Figura 24: Tela principal do WEKA.

O projeto inclui algoritmos para regressão, classificação, agrupamento, mineração de regras de associação e seleção de atributo. A exploração preliminar de dados é bem atendida por telas de visualização de dados e muitas ferramentas de pré-processamento. Estas, quando combinados com avaliação estatística de esquemas de aprendizagem e visualização dos resultados de aprendizagem, suporta modelos de processo de mineração de dados.

O WEKA conta também com um gerenciador de pacotes (vide figura 25), com o qual é possível pesquisar e baixar diversas extensões feitas pela comunidade.

Dentre elas encontram-se algoritmos de pré-processamento, seleção de características, classificação, regressão etc. Também é possível instalar pacotes não-oficiais, que muitas vezes ainda estão sendo desenvolvidos e, por isso, ainda não foram disponibilizados oficialmente.

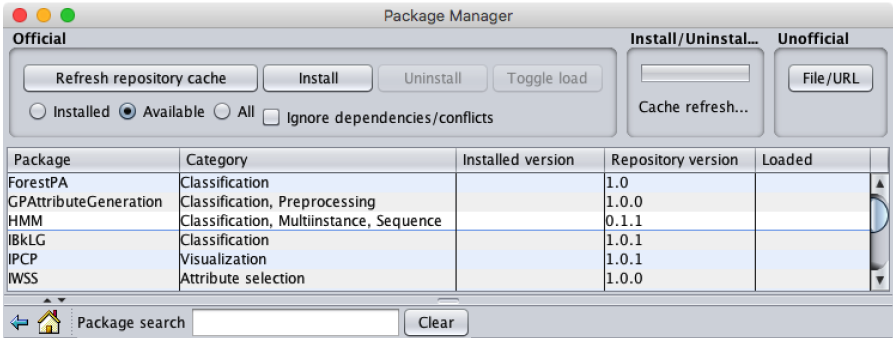


Figura 25: Tela do gerenciador de pacotes do WEKA.

Todos os oito métodos de classificação propostos neste trabalho estão disponíveis no WEKA. Nos experimentos com máquinas de vetores de suporte, as célebres bibliotecas LIBLINEAR (FAN et al., 2008) e LIBSVM (CHANG; LIN, 2011), totalmente suportadas pelo WEKA, poderiam ter sido usadas para treinar as máquinas de vetores de suporte linear e polinomial, respectivamente.

Na página da LIBSVM, porém, é possível encontrar uma versão alternativa da LIBLINEAR, chamada LIBLINEAR-poly2, cujo código "é mais eficiente do que a LIBSVM em ambas as fases de treinamento e teste para SVM com *kernels* polinomiais de grau 2", como afirmam Chang et al. (2010) na documentação da biblioteca. Neste mesmo documento, há um exemplo mostrando que a diferença de precisão de classificação entre elas foi irrisória (da ordem de 0,001%) e que o tempo de execução foi cerca de sete vezes menor na segunda biblioteca.

Portanto, a fim de manter a comparação entre SVMs consistente em termos de plataforma/linguagem e, ao mesmo tempo, tirar proveito das vantagens da LIBLINEAR-poly2, decidiu-se por adotar as bibliotecas LIBLINEAR e LIBLINEAR-poly2 para treinar as SVMs linear e polinomial, respectivamente.

4.3 Conjuntos de dados utilizados

O treinamento dos algoritmos de aprendizado de máquina exige conjuntos de dados que sejam suficientemente representativos para o problema proposto. Os estudos avaliativos dos classificadores utilizaram três bases de dados públicas e uma privada, compostas por *e-mails* reais. A descrição dos *corpora* é dada a seguir.

4.3.1 Ling Spam

Compilado e disponibilizado em domínio público por Androutsopoulos et al. (2000a), o *corpus* Ling Spam (ANDROUTSOPOULOS et al., 2000b) é composto por 2893 mensagens de *e-mail* e apresenta as seguintes características:

- 2.412 *hams* de uma lista de discussão moderada sobre a profissão e ciência de linguística, obtidas por meio do *download* aleatório de resumos dos arquivos, separação de mensagens e remoção de texto adicionado pelo servidor da lista.
- 481 *spams* recebidos pelo primeiro autor. Anexos, *tags* HTML e mensagens duplicadas recebidas no mesmo dia não foram incluídos.
- É disponibilizado em quatro versões:
  - *bare*: lematização desativada e remoção de palavras vazias desativada.
  - *lemm*: lematização ativada e remoção de palavras vazias desativada.
  - *lemm_stop*: lematização ativada e remoção de palavras vazias ativada.
  - *stop*: lematização desativada e remoção de palavras vazias ativada.

Embora as mensagens de *spam* e *ham* sejam extraídas de fontes diferentes e algumas informações (ex: anexos e *tags* HTML) tenham sido removidas, este conjunto apresentou uma avaliação mais realista do que qualquer outro disponível quando do seu lançamento, sendo usado em diversos estudos, como Barigou, Beldjilali e Atmani (2014), Kaya e Ertuğrul (2016), Shams e Mercer (2016), Yang et al. (2016) e Kumaresan, Saravanakumar e Balamurugan (2017).

Neste trabalho, as quatro versões foram consolidadas em um único *corpus*.

4.3.2 Spam Assassin

O *corpus* Spam Assassin (The Apache Software Foundation, 2005) é composto por 6.047 mensagens de *e-mail* reais, com aproximadamente 31% de *spam*. As mensagens provêm de várias fontes, incluindo fóruns públicos e mensagens enviadas de e para os desenvolvedores do Spam Assassin. Todos os cabeçalhos foram reproduzidos em sua totalidade. Há ofuscação de endereços[3] e os *hostnames* foram trocados por `spamassassin.taint.org` em algumas situações. Entretanto, na maioria dos casos, os cabeçalhos aparecem da mesma forma que foram recebidos. O *corpus* é composto por cinco partes, descritas na tabela 5.

Tabela 5: Subconjuntos do *corpus* Spam Assassin.

Subconjunto	*Hams*	*Spams*	Descrição
spam	-	500	*Spams* recebidos de fontes reais
easy_ham	2.500	-	Mensagens de fácil identificação, por serem completamente diferentes de spams
hard_ham	250	-	Mensagens de difícil identificação, por se parecerem em muitos aspectos com o spam típico
easy_ham_2	1.400	-	Adições recentes ao *corpus*
spam_2	-	1.397	

O *corpus* tem sido usado tanto para o desenvolvimento do *software* Apache Spam Assassin (FOUNDATION, 2015) e outros filtros de *spam* quanto como alvo de avaliação em publicações como Meyer e Whateley (2004), Carpinteiro et al. (2006), Sirisanyalak e Sornil (2007), Pérez-Díaz et al. (2012), Barigou, Beldjilali e Atmani (2014), Kaya e Ertuğrul (2016) e Shams e Mercer (2016).

Neste trabalho, as cinco partes deste *corpus* foram consolidadas, gerando um conjunto com mais mensagens e maior variabilidade.

[3]Técnica que visa esconder (ou dificultar) a identificação de *e-mail* por *bots*, mantendo-os ao mesmo tempo visíveis para humanos.

4.3.3 TREC

O TREC Spam Track (CORMACK; LYNAM, 2005) (CORMACK; LYNAM, 2006) (CORMACK; LYNAM, 2007) foi a maior e mais realista avaliação de laboratório até o momento de sua disponibilização. Nos três anos que funcionou, dez conjuntos de teste com um total de 721.461 mensagens foram utilizados para testar filtros enviados por 35 participantes. Cada participante apresentou até quatro filtros, que foram testados com os *e-mails* dos *corpora*. (CORMACK et al., 2008)

Dentre os diversos *corpora* disponíveis, três foram utilizados neste trabalho:

- *TREC 2005* (CORMACK; LYNAM, 2005): é composto pelo *e-mail* armazenado de 150 executivos da Enron[4], coletados e lançados no domínio público como resultado da investigação federal dos EUA sobre seu colapso. Essas mensagens foram rotuladas e somadas a mensagens de *spam* oriundas de uma fonte pública, alteradas de forma a parecer terem sido entregues à Enron durante o mesmo período. Esta base estabeleceu que os *corpora* privados e públicos poderiam produzir resultados razoavelmente consistentes.
- *TREC 2006* (CORMACK; LYNAM, 2006): compreende dois *corpora* privados e dois públicos (*trec06p* - em inglês, extraído da Web e acrescido de *spam* público, e e *trec06c* - em chinês, obtido a partir de uma lista de *e-mails* e acrescido de *spams* capturados no mesmo site por um *honeypot*[5]).
- *TREC 2007* (CORMACK; LYNAM, 2007): compreende um *corpus* privado e um público (*trec07p*), sendo o último composto por mensagens entregues a um servidor particular durante três meses.

Neste trabalho, os *corpora* públicos de 2005, 2006 (somente a versão em inglês) e 2007 foram consolidados, gerando um conjunto de origem mista e com uma quantidade expressiva de mensagens.

[4]Companhia de energia americana que decretou falência em 2001 e foi alvo de investigação federal após ter sido descoberto um escândalo fiscal da ordem de bilhões de dólares.

[5]Campo invisível inserido em formulários de páginas da Web, que são preenchidos somente por *bots* e servem como evidência da natureza não humana e maliciosa destes mecanismos.

4.3.4 Unifei 2017

A base de dados Unifei foi coletada durante o segundo semestre de 2016 na Universidade Federal de Itajubá (Unifei) pelo Grupo de Pesquisas em Engenharia de Sistemas e de Computação (GPESC). Consiste em um universo de *e-mails* que representa a realidade da universidade, com mensagens recebidas pelos professores, STAs e alunos da universidade.

Contendo 862.229 *e-mails*, sendo 353.152 *hams* e 509.077 *spams*, é um *corpus* consideravelmente maior que os outros já apresentados. Entretanto, é importante mencionar que os rótulos dos *e-mails* não foram atribuídos manualmente, mas sim de forma automática pelo anti-*spam* proprietário CanIt-PRO da Roaring Penguin.

Cada mensagem foi submetida aos seguintes pré-processamentos:

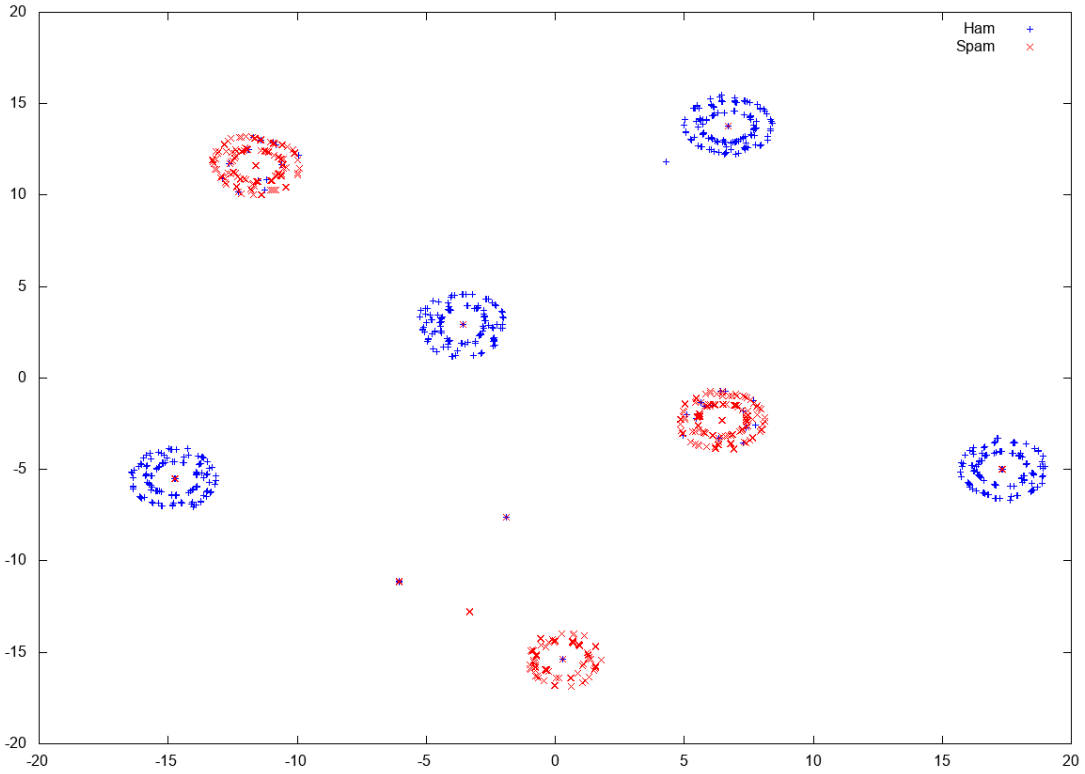
- Substituição de *tags* HTML (e seus atributos), símbolos monetários, *links*, números, palavras pequenas[6] e grandes[7] por *tokens*.
- Remoção de pontuação, substituição de caracteres com diacríticos[8] e/ou maiúsculos por seus equivalentes minúsculos simples e correção de codificação.

Depois de pré-processar as mensagens, foram executados os métodos de seleção de características (CHI2, FD e MI) de forma a obter representações vetoriais (numéricas) de cada mensagem.

Por fim, utilizando a técnica t-SNE (do inglês *t-Distributed Stochastic Neighbor Embedding*) (MAATEN; HINTON, 2008), foram geradas visualizações bidimensionais do conjunto para cada método de seleção de características, a fim de detectar possíveis inconsistências nos dados. Os resultados apresentados nas figuras 26, 27 e 28, indicam a existência de amostras ambíguas (ou seja, mensagens de classes diferentes mas que ocupam a mesma posição no gráfico). Tal problema motivou a criação do conjunto Unifei 2018, que será apresentado a seguir, na seção 4.3.5.

[6] Cuja quantidade de caracteres é inferior a um limite pré-estabelecido.
[7] Cuja quantidade de caracteres é superior a um limite pré-estabelecido.
[8] Sinais gráficos que alteram a realização fonética (som) de uma letra.

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

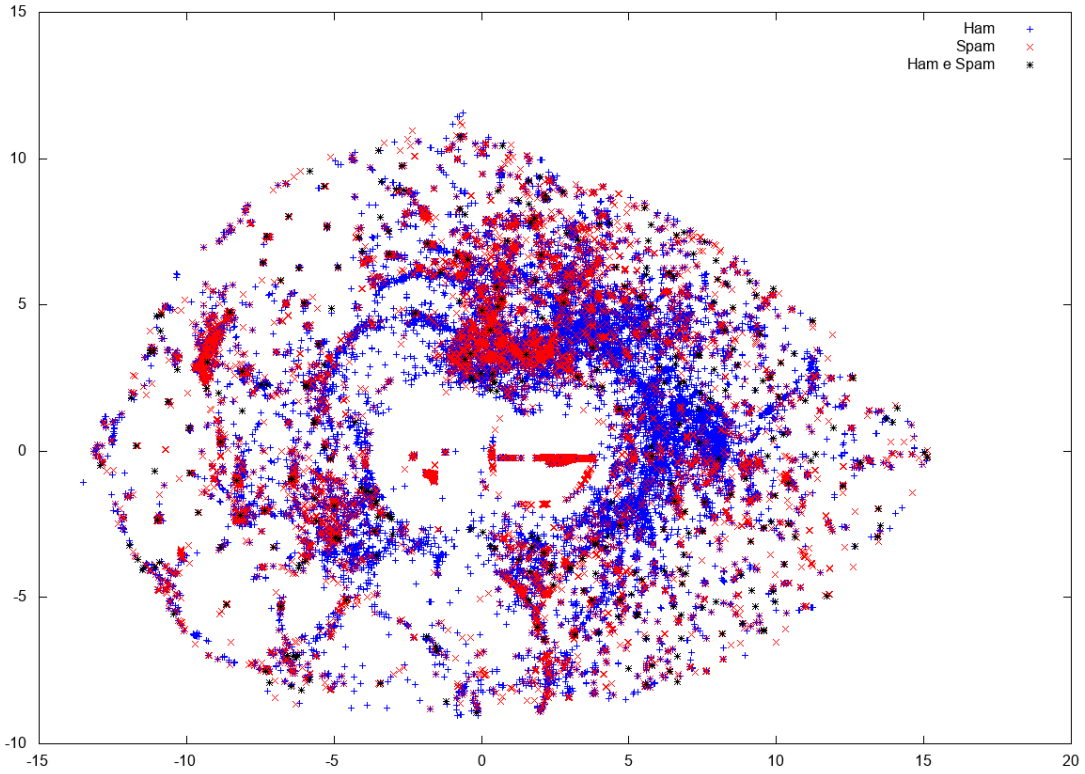
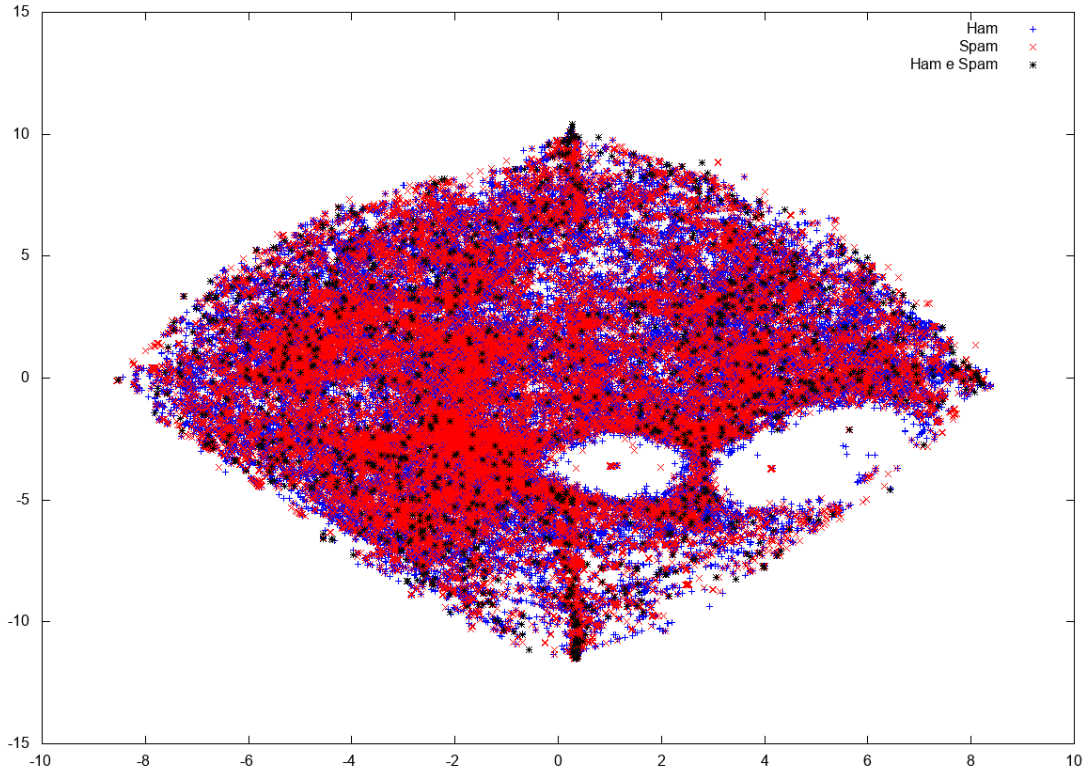


Figura 26: Visualização 2D da base Unifei 2017 (método CHI2).

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

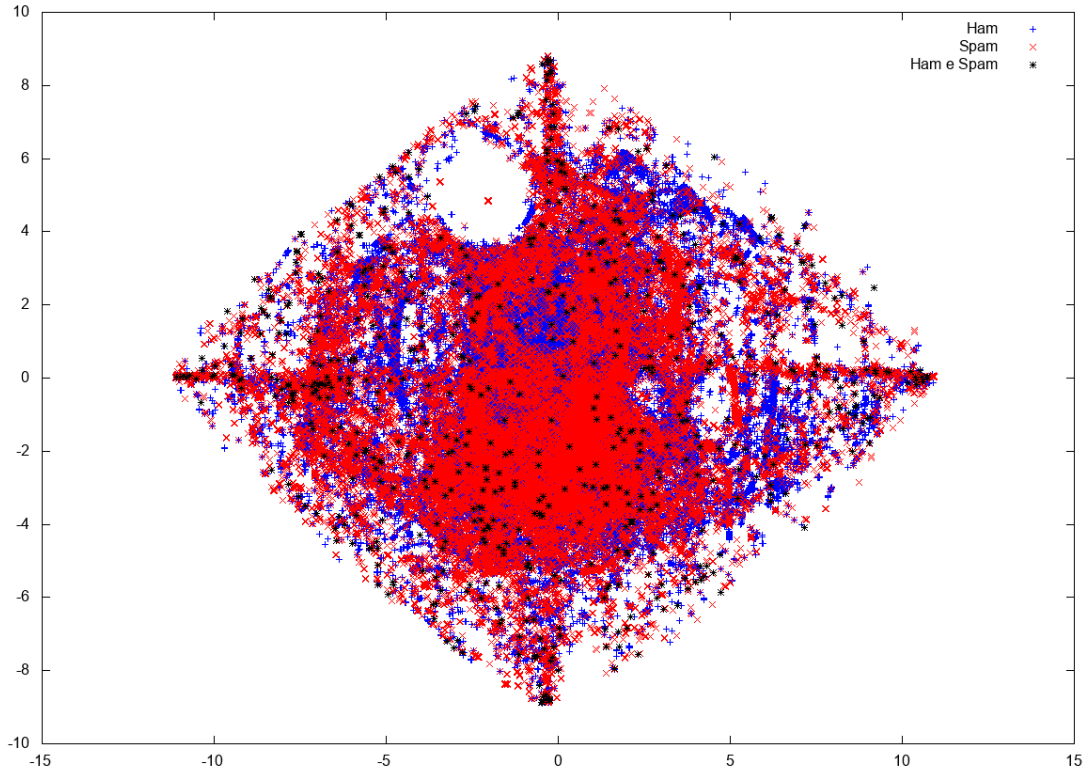
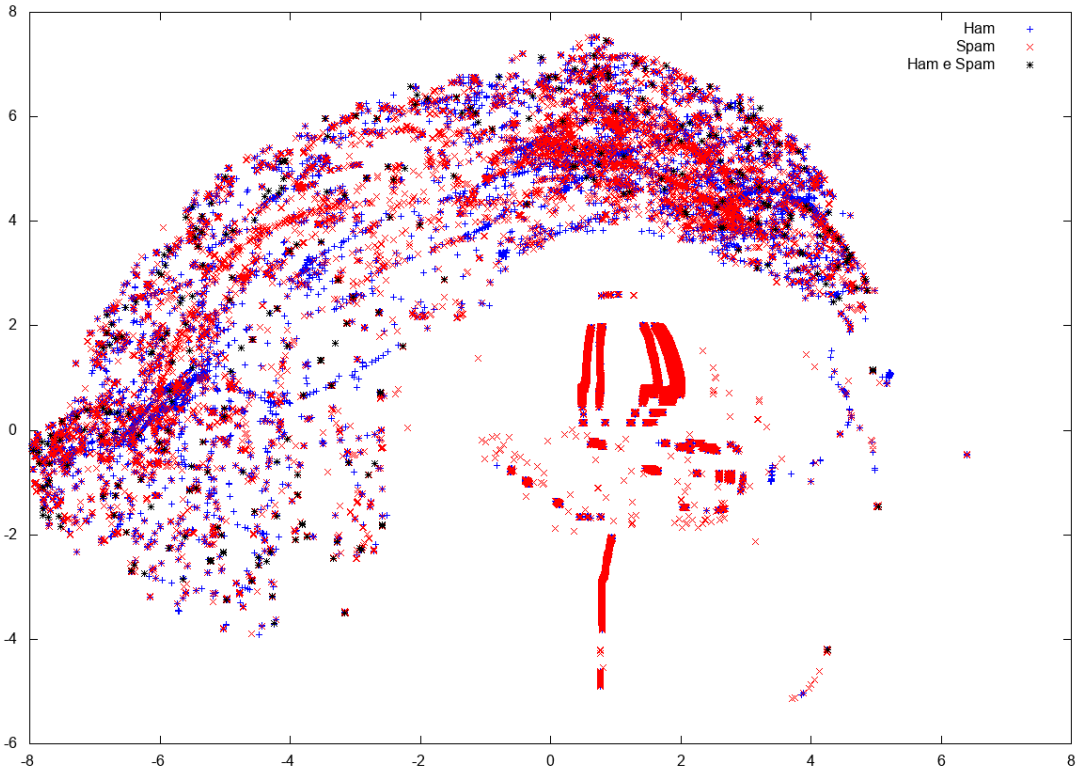


Figura 27: Visualização 2D da base Unifei 2017 (método FD).

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

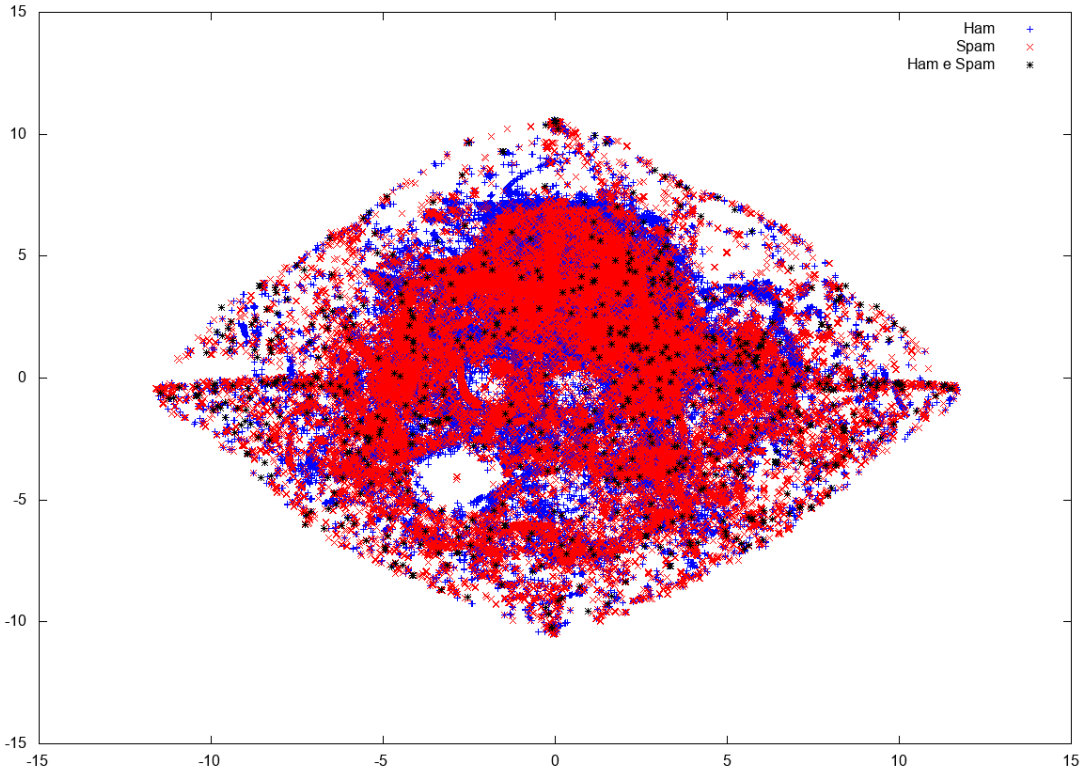


Figura 28: Visualização 2D da base Unifei 2017 (método MI).

4.3.5 Unifei 2018

Considerando que alguns *e-mails* (potencialmente de classes distintas) diferem-se totalmente de todos os outros do conjunto e, portanto, não apresentam qualquer relação de informação mútua nem possuem as palavras mais frequentes do conjunto, as representações vetoriais destas mensagem tornam-se vazias. Neste sentido, os dois problemas a seguir foram apontados na base Unifei 2017:

1. Existência de *e-mails* de classes diferentes com representações vetoriais iguais;
2. Presença de vetores nulos (ou seja, $\vec{x} = [x_1, x_2, \ldots, x_n] = [0, 0, \ldots, 0]$).

Visando endereçar as consequências do primeiro problema, foi realizado um procedimento de rerrotulagem de mensagens/vetores. Considerando que haja H *hams* e S *spams* com representação vetorial igual, o procedimento resume-se a:

- Se $H > S$, então os S *spams* são rerrotulados como *hams*.
- Se $S > H$, então os H *hams* são rerrotulados como *spams*.

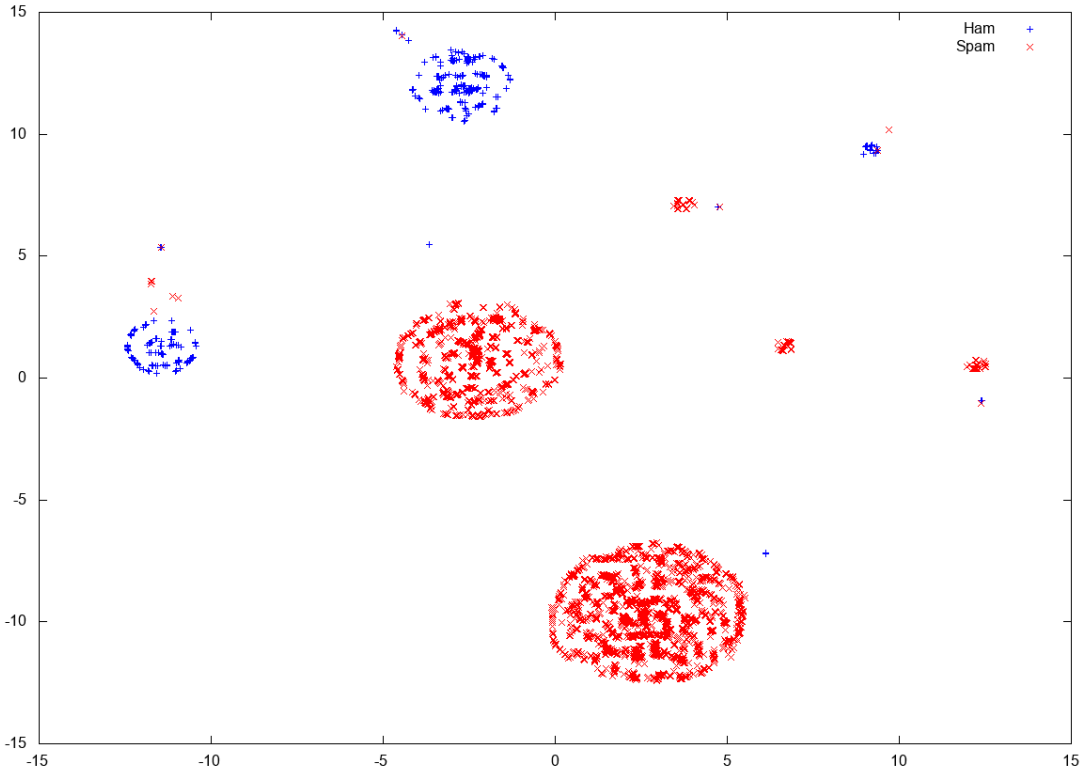
Para contornar o segundo problema, os vetores nulos foram incluídos apenas na etapa de testes (classificação), ou seja, eles não foram apresentados aos métodos de aprendizado de máquina durante a etapa de treinamento (geração de modelo).

Portanto, a base Unifei 2018 consiste em uma versão corrigida e melhorada da base Unifei 2017, obtida pela execução dos seguintes passos:

1. Uso de todos os *e-mails* da base Unifei 2017;
2. Detecção de **mensagens** duplicadas (de classes diferentes);
3. Rerrotulagem das **mensagens** duplicadas identificadas na etapa anterior;
4. Execução dos métodos de seleção de características para obtenção dos vetores (semelhante à Unifei 2017);
5. Detecção dos **vetores** duplicados (de classes diferentes);
6. Rerrotulagem dos **vetores** duplicados identificadas na etapa anterior.

Os gráficos das figuras 29, 30 e 31 comprovam a ausência de amostras ambíguas (ou seja, mensagens de classes diferentes que ocupam a mesma posição no gráfico).

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

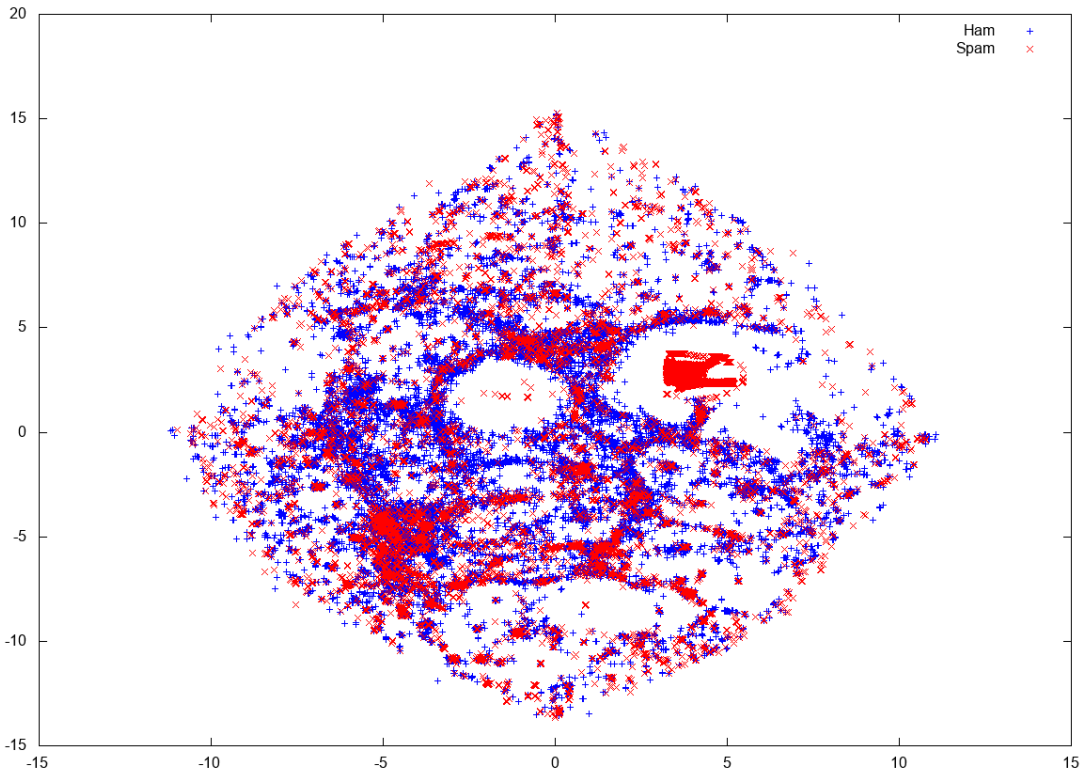
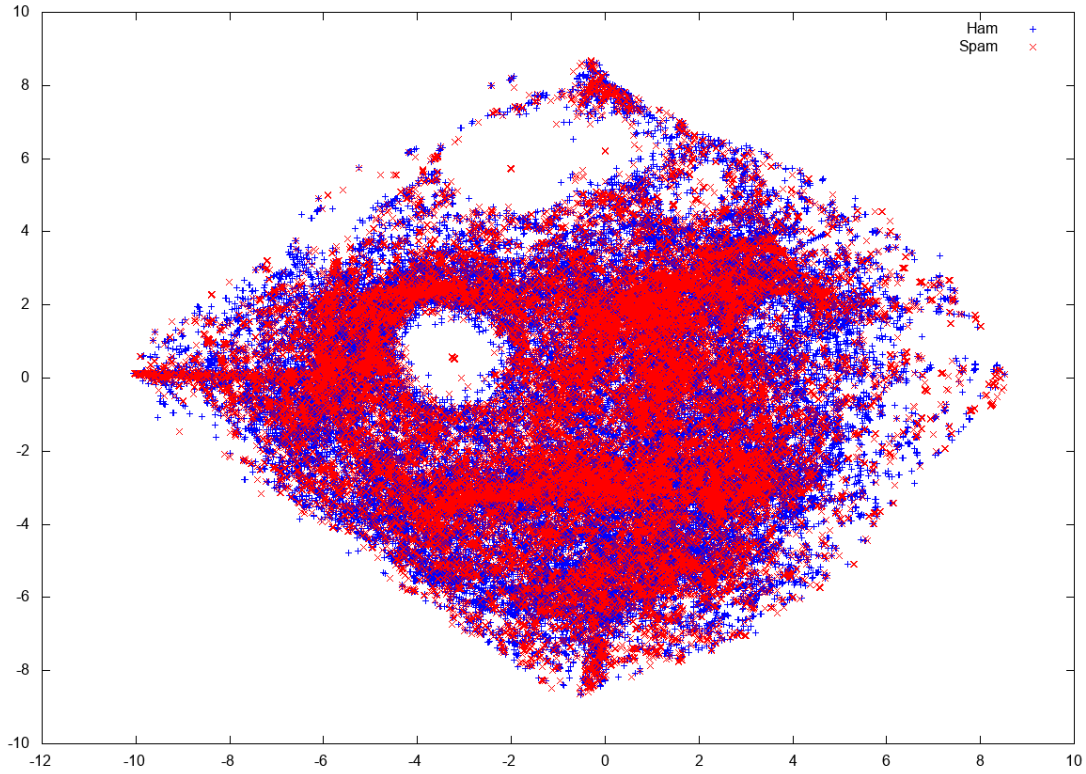


Figura 29: Visualização 2D da base Unifei 2018 (método CHI2).

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

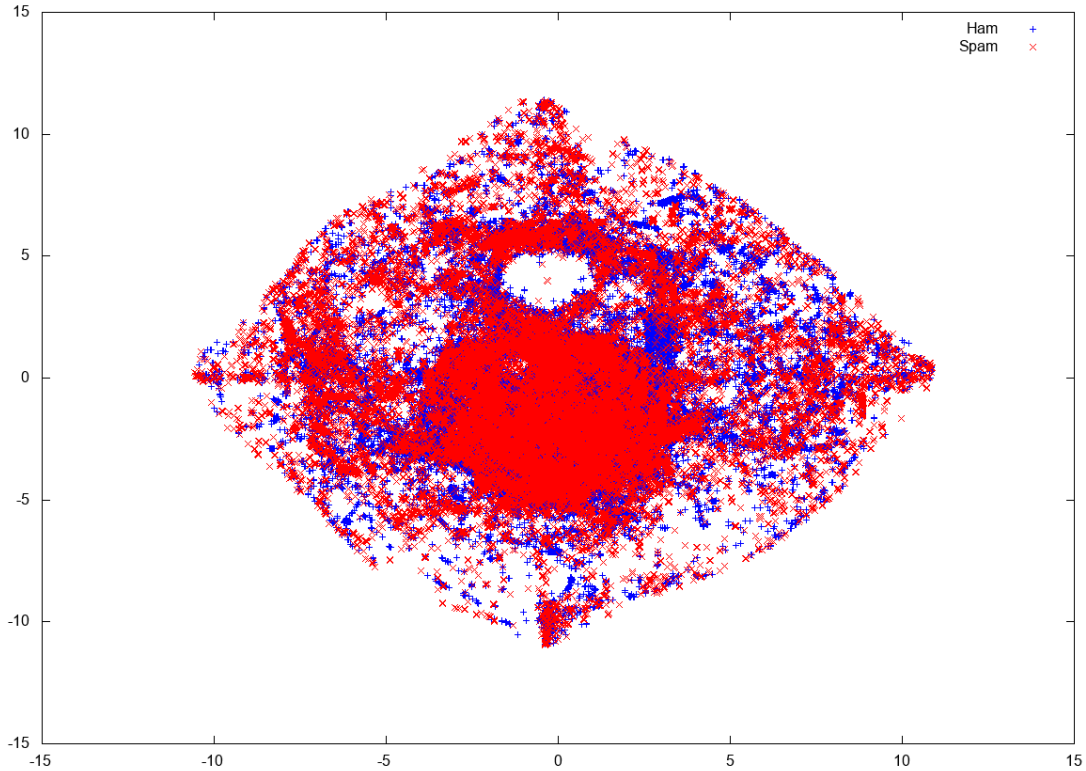
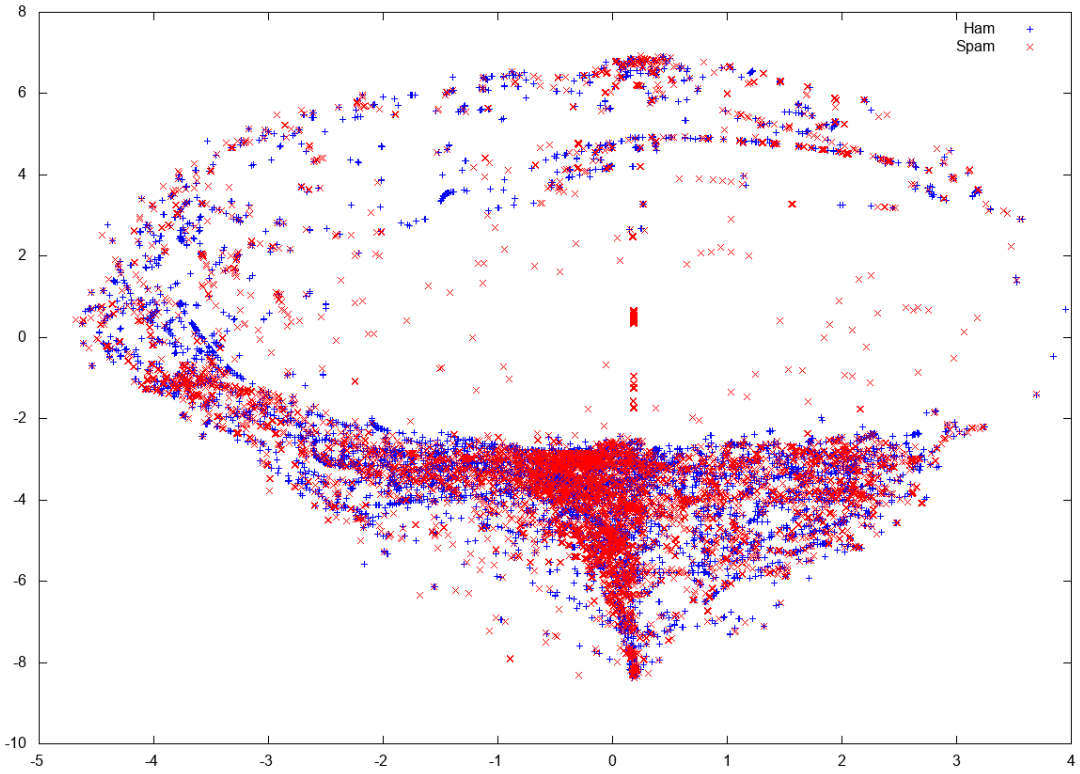


Figura 30: Visualização 2D da base Unifei 2018 (método FD).

(a) Resultado do t-SNE com vetores de 8 características.



(b) Resultado do t-SNE com vetores de 1024 características.

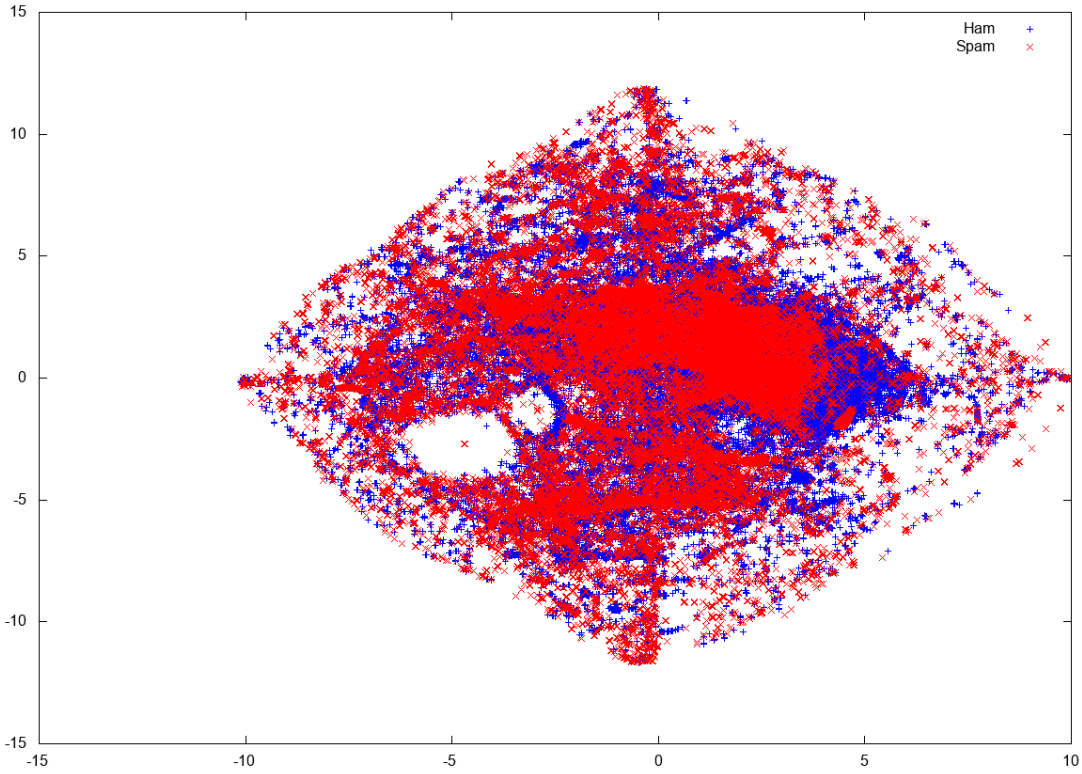


Figura 31: Visualização 2D da base Unifei 2018 (método MI).

4.4 Métricas de desempenho

Trabalhos que envolvam comparações de qualquer natureza requerem a definição de qual é o sujeito de avaliação, bem como sob quais óticas este será avaliado.

No contexto de sistemas anti-*spam*, é desejável que estes sistema não classifiquem mensagens legítimas incorretamente, caso contrário o usuário poderá perder conteúdo importante de forma irreversível, dependendo da política de tratamento de *spam* do sistema. Quando o sistema falha neste tipo de classificação, diz-se que foi obtido um resultado falso positivo.

Além disso, também deseja-se que estes sejam capazes de identificar e filtrar a maior quantidade possível de mensagens indesejadas, protegendo o destinatário de conteúdo potencialmente malicioso ou que implique em perda financeira e/ou de produtividade. Quando o sistema falha neste tipo de classificação, diz-se que foi obtido um resultado falso negativo.

Tendo estas definições em mente, são propostas as métricas de precisão e revocação para avaliação dos modelos de aprendizado de máquina, que correspondem às ausências de resultados falsos-positivos e falsos-negativos, respectivamente. Nos apêndices, os resultados são dados em termos de *ham* e *spam* separadamente.

Adicionalmente, o escore F_1 (explicado em detalhes na seção 4.4.3) faz uma combinação da precisão e revocação, apresentando-se como uma métrica capaz de avaliar um modelo em termos de ambos os tipos de erro e servindo como medida geral da eficácia de um sistema anti-*spam*, já que os resultados são dados em termos de *ham* e *spam* simultaneamente.

Por fim, são apresentadas métricas relacionadas aos tempos de treinamento e execução. A primeira mostra o tempo necessário para construir um modelo preditivo a partir das amostras de treinamento, e a segunda consiste no intervalo necessário para classificar todo o conjunto de testes usando o modelo em questão.

Vale destacar que os resultados são dados em termos de média e intervalo de confiança (com $C = 0,95$) (NEYMAN, 1937) considerando todas as dez execuções.

Para as equações (4.1), (4.2), (4.3), (4.4), (4.5), (4.6), (4.7), (4.8) e (4.9), considere que:

- N_{HAM} é a quantidade total de *hams* no conjunto de testes;
- N_{SPAM} é a quantidade total de *spams* no conjunto de testes;
- $n_{H \to H}$ é a quantidade de *hams* classificados corretamente;
- $n_{H \to S}$ é a quantidade de *hams* classificados como *spam*;
- $n_{S \to S}$ é a quantidade de *spams* classificados corretamente;
- $n_{S \to H}$ é a quantidade de *spams* classificados como *ham*.

4.4.1 Precisão

Pode ser vista como uma medida de exatidão, pois denota a ausência de falsos positivos. É calculada tanto para *ham* quanto para *spam* pelas equações (4.1) e (4.2), respectivamente. A precisão geral é dada pela equação (4.3).

$$P(HAM) = \frac{n_{H \to H}}{n_{H \to H} + n_{S \to H}} \quad (4.1) \qquad P(SPAM) = \frac{n_{S \to S}}{n_{S \to S} + n_{H \to S}} \quad (4.2)$$

$$P(GERAL) = \frac{N_{HAM} * P(HAM) + N_{SPAM} * P(SPAM)}{N_{HAM} + N_{SPAM}} \quad (4.3)$$

4.4.2 Revocação

Pode ser vista como uma medida de completude, pois denota a ausência de falsos negativos. É calculada tanto para *ham* quanto para *spam* pelas equações (4.4) e (4.5), respectivamente. A revocação geral é dada pela equação (4.6).

$$R(HAM) = \frac{n_{H \to H}}{n_{H \to H} + n_{H \to S}} \quad (4.4) \qquad R(SPAM) = \frac{n_{S \to S}}{n_{S \to S} + n_{S \to H}} \quad (4.5)$$

$$R(GERAL) = \frac{N_{HAM} * R(HAM) + N_{SPAM} * R(SPAM)}{N_{HAM} + N_{SPAM}} \quad (4.6)$$

4.4.3 Escore F_1

Na análise estatística da classificação binária, o escore F_1 é uma medida que considera tanto a precisão como a revocação de um teste para calcular sua pontuação. É dada pela média harmônica entre estes valores (vide equações (4.7), (4.8) e (4.9)) e atinge seu melhor e pior valor em 1 e 0, respectivamente.

$$F_1(HAM) = 2 \cdot \frac{1}{\frac{1}{P(HAM)} + \frac{1}{R(HAM)}} = 2 \cdot \frac{P(HAM) \cdot R(HAM)}{P(HAM) + R(HAM)} \tag{4.7}$$

$$F_1(SPAM) = 2 \cdot \frac{1}{\frac{1}{P(SPAM)} + \frac{1}{R(SPAM)}} = 2 \cdot \frac{P(SPAM) \cdot R(SPAM)}{P(SPAM) + R(SPAM)} \tag{4.8}$$

$$F_1(GERAL) = \frac{N_{HAM} * F_1(HAM) + N_{SPAM} * F_1(SPAM)}{N_{HAM} + N_{SPAM}} \tag{4.9}$$

4.4.4 Tempo de Treinamento

É definido como o tempo que o algoritmo de classificação levou para ser treinado e gerar um modelo aplicável ao conjunto de testes. É dado por:

$$t_{treinamento} = t'_T - t_T \tag{4.10}$$

Em que t'_T e t_T marcam, respectivamente, os instantes inicial e final do treinamento.

4.4.5 Tempo de Classificação

É definido como o tempo que o modelo levou para classificar todas as amostras do conjunto de treinamento. É dado por:

$$t_{classificação} = t'_C - t_C \tag{4.11}$$

Em que t'_C e t_C marcam, respectivamente, os instantes inicial e final da classificação.

5 Experimentos e Resultados

Visando facilitar a análise e entendimento dos resultados, as conclusões foram agrupadas em três seções que representam cada aspecto macroscópico deste trabalho:

- Seleção de características e padrões zerados (seção 5.1)
  - métodos de seleção de características e quantidade de padrões zerados;
  - quantidade de características e quantidade de padrões zerados;
  - métodos de seleção de características e desempenho (em termos de escore F_1 médio) em cada base de dados.
- Redução de dimensionalidade (seção 5.2)
  - métodos de seleção de características e quantidades (absoluta e relativa) de características após a redução de dimensionalidade;
  - proporcionalidade entre as quantidades de características antes e após a redução de dimensionalidade, bem como as anomalias identificadas.
- Treinamento e classificação (seção 5.3)
  - comparação dos escores F_1, tempo de treinamento e de classificação obtidos pelos métodos de classificação no *ranking* de cada base de dados;
  - métodos de seleção de características e ocorrências no *ranking*;
  - quantidade original de características e ocorrências no *ranking*.

Vale ressaltar que os resultados experimentais, usados para *renderizar* as visualizações e tirar as conclusões, foram incluídos na íntegra nos apêndices. Além disso, detalhes sobre as configurações do *hardware* e as versões específicas dos *softwares* utilizados também foram disponibilizados.

5.1 Seleção de características e padrões zerados

A tabela 6 e as figuras 32, 33, 34, 35, 36 e 37 mostram, para cada método de seleção de características, os percentuais de padrões zerados em cada conjunto de dados.

Tabela 6: Quantidades de padrões zerados nos conjuntos de dados — NF: número de características; H_0: percentual de *hams* zerados; S_0: percentual de *spams* zerados.

Método	NF	Ling Spam		Spam Assassin		TREC		Unifei 2017		Unifei 2018	
		$H_0(\%)$	$S_0(\%)$	$H_0(\%)$	$S_0(\%)$	$H_0(\%)$	$S_0(\%)$	$H_0(\%)$	$S_0(\%)$	$H_0(\%)$	$S_0(\%)$
CHI2	8	97.8	99.3	4.2	11.7	99.6	98.9	99.7	99.9	98.8	98.6
	16	95.8	99.3	3.9	9.9	99.1	97.5	98.6	98.9	98.0	97.6
	32	91.9	99.3	3.5	9.1	98.9	97.2	98.4	98.8	96.0	96.2
	64	85.2	99.3	2.6	8.5	98.2	96.6	98.3	98.8	93.6	94.6
	128	74.2	99.3	2.0	6.5	97.2	94.6	97.8	98.7	89.9	88.7
	256	59.7	99.3	0.6	0.6	94.1	91.9	95.5	97.9	85.9	86.9
	512	34.2	61.9	0.6	0.6	82.2	78.4	86.3	88.2	69.5	60.2
	1024	34.2	49.1	0.5	0.5	57.5	45.1	42.2	47.3	41.2	36.2
FD	8	0.0	0.0	0.0	0.7	4.4	2.4	0.8	2.1	0.2	2.5
	16	0.0	0.0	0.0	0.2	4.3	2.4	0.8	2.1	0.1	2.5
	32	0.0	0.0	0.0	0.2	4.2	2.4	0.8	2.1	0.1	2.5
	64	0.0	0.0	0.0	0.2	4.1	2.4	0.8	2.1	0.1	2.5
	128	0.0	0.0	0.0	0.2	4.0	2.4	0.8	2.0	0.1	2.5
	256	0.0	0.0	0.0	0.2	4.0	2.4	0.8	2.0	0.1	2.5
	512	0.0	0.0	0.0	0.2	3.9	2.3	0.8	2.0	0.1	2.5
	1024	0.0	0.0	0.0	0.2	3.9	2.3	0.7	2.0	0.1	2.5
MI	8	0.0	0.0	0.0	1.0	7.8	3.0	15.1	14.9	14.9	15.0
	16	0.0	0.0	0.0	0.9	5.9	3.0	4.7	6.7	4.0	7.2
	32	0.0	0.0	0.0	0.2	4.4	2.8	4.7	6.7	4.0	7.2
	64	0.0	0.0	0.0	0.2	4.3	2.4	0.8	2.1	0.1	2.5
	128	0.0	0.0	0.0	0.2	4.3	2.4	0.8	2.1	0.1	2.5
	256	0.0	0.0	0.0	0.2	4.2	2.4	0.8	2.0	0.1	2.5
	512	0.0	0.0	0.0	0.2	4.2	2.4	0.8	2.0	0.1	2.5
	1024	0.0	0.0	0.0	0.2	3.9	2.3	0.8	2.0	0.1	2.5

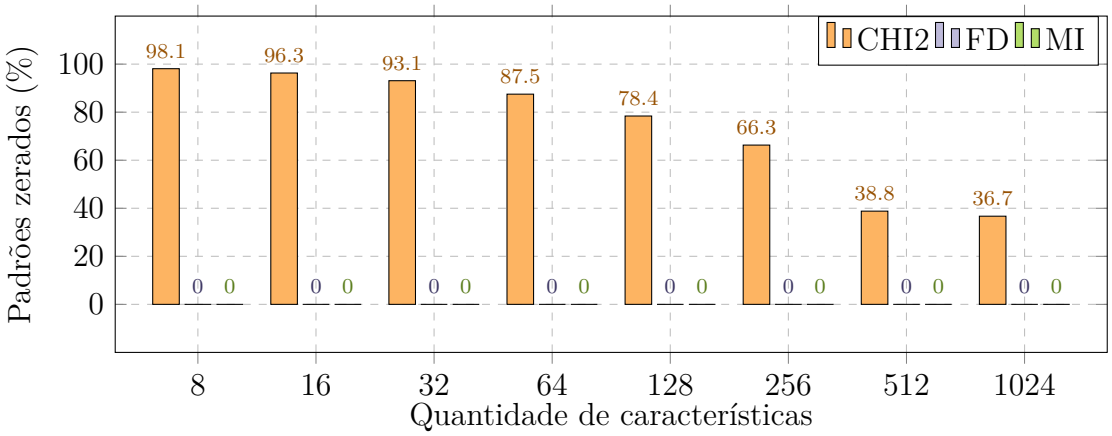


Figura 32: Padrões zerados na base Ling Spam.

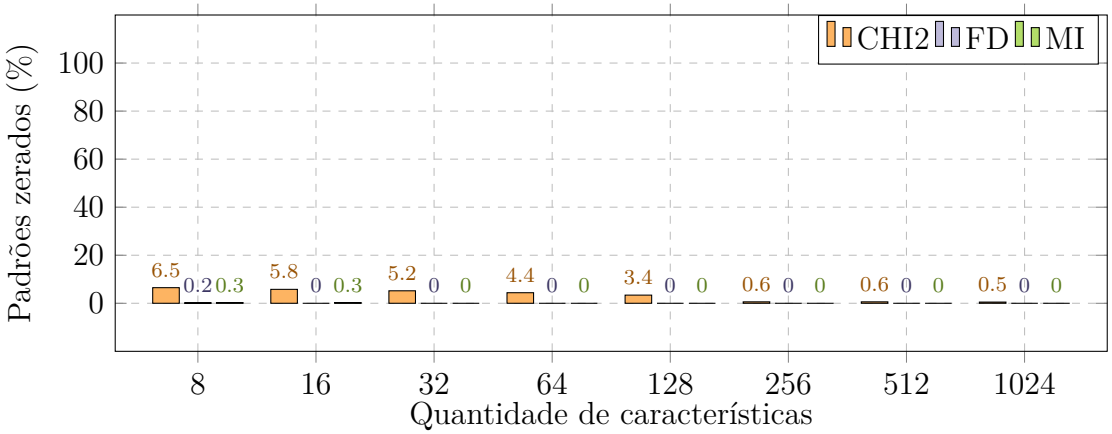


Figura 33: Padrões zerados na base Spam Assassin.

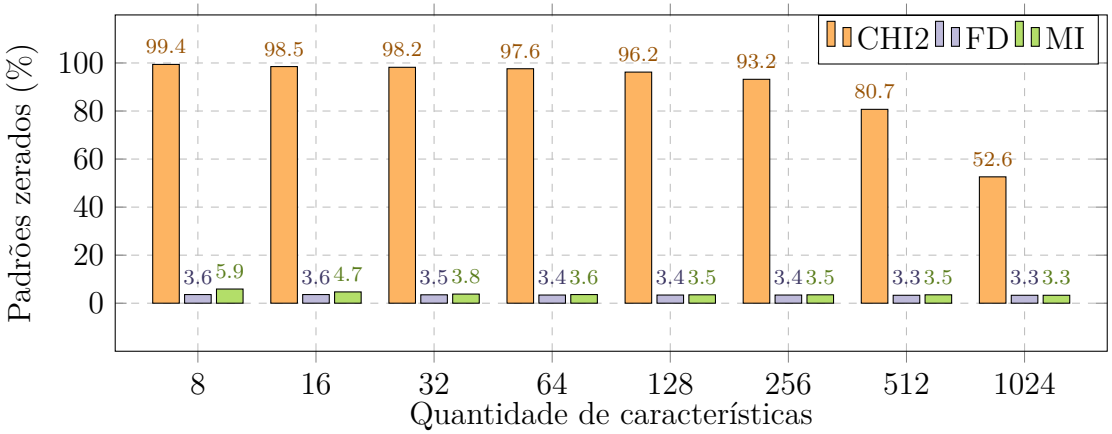


Figura 34: Padrões zerados na base TREC.

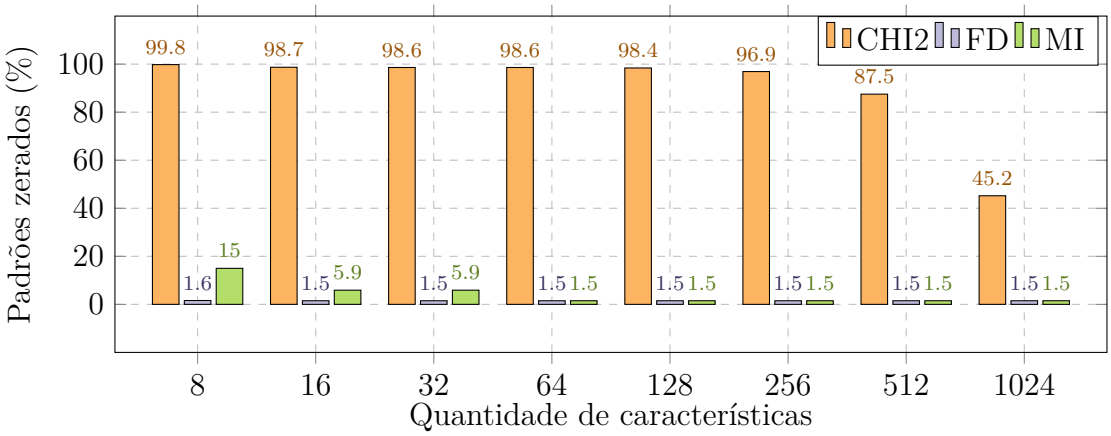


Figura 35: Padrões zerados na base Unifei 2017.

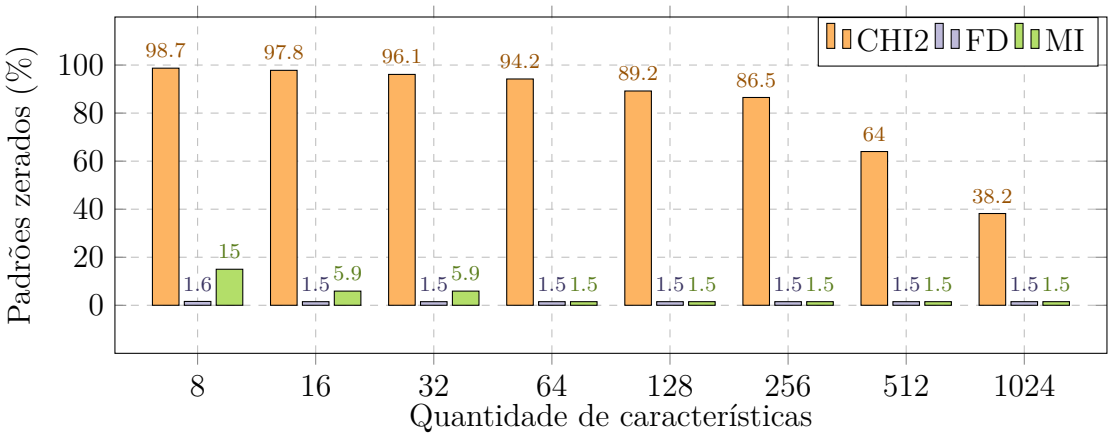


Figura 36: Padrões zerados na base Unifei 2018.

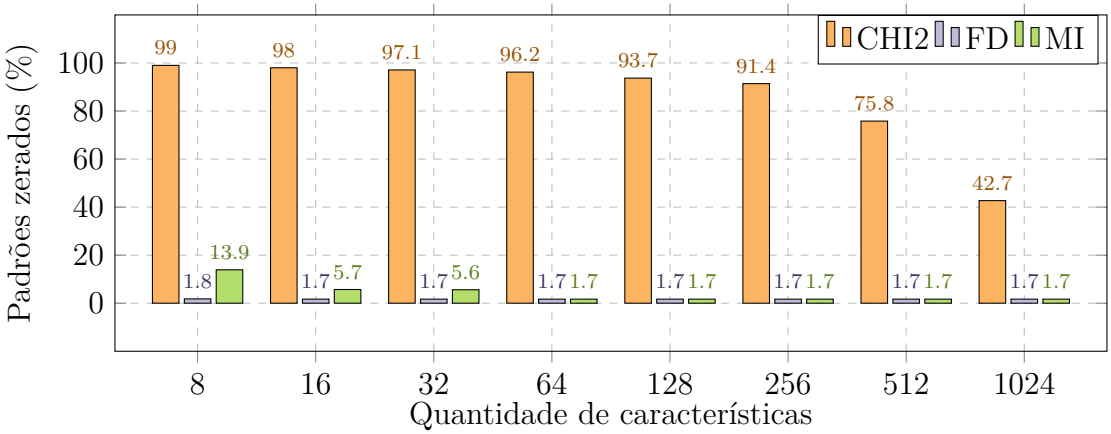


Figura 37: Padrões zerados em todas as bases.

Para cada conjunto de dados, os resultados de todos os modelos de aprendizado de máquina foram agrupados em faixas de escore F_1 médio — "ruim" (<80%), "médio" (entre 80% e 90%) ou "bom" (>90%). Em seguida, foram contabilizadas as ocorrência dos métodos de seleção de características em cada faixa de escore F_1. Estes resultados estão representados nas figuras 38, 39, 40, 41, 42 e 43.

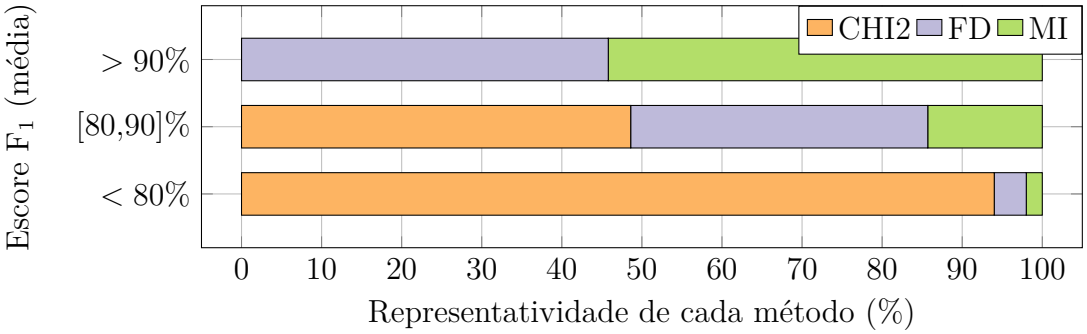


Figura 38: Escore F_1 vs métodos de seleção de características (Ling Spam).

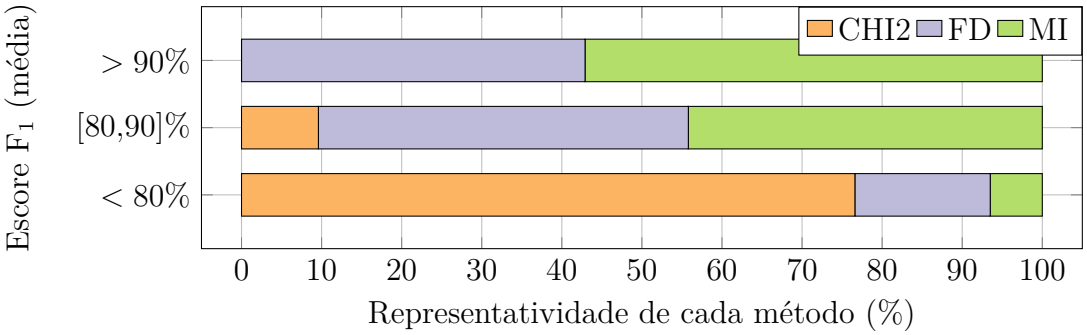


Figura 39: Escore F_1 vs métodos de seleção de características (Spam Assassin).

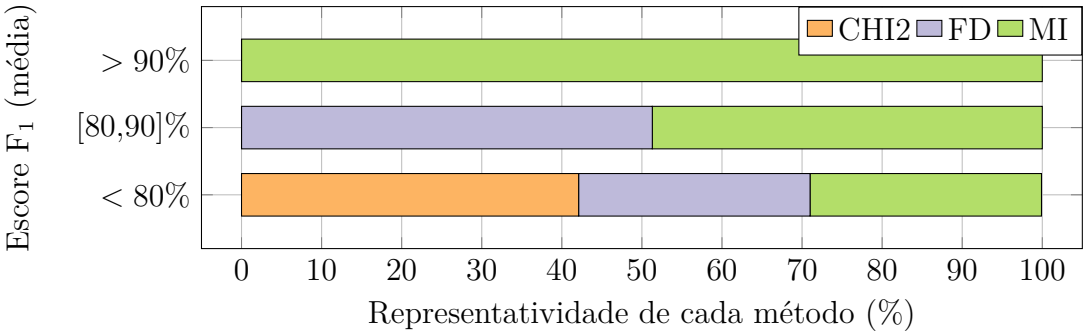


Figura 40: Escore F_1 vs métodos de seleção de características (TREC).

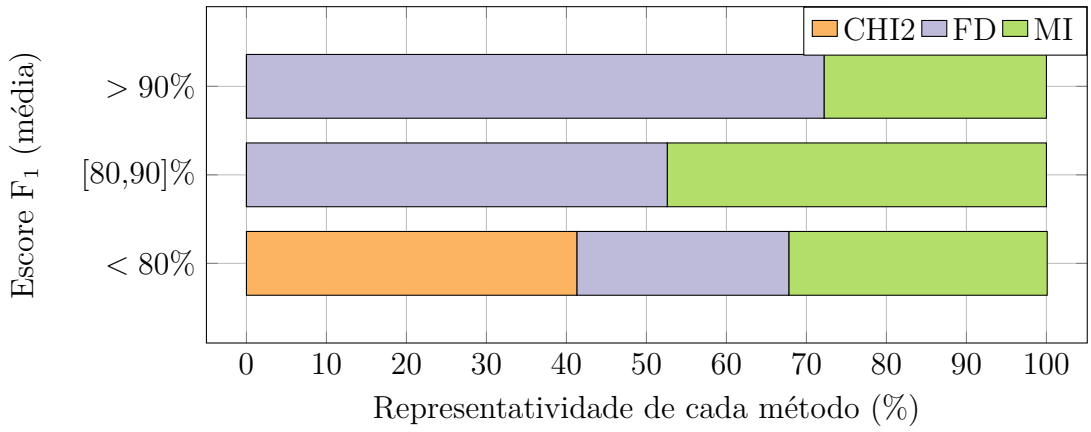


Figura 41: Escore F_1 vs métodos de seleção de características (Unifei 2017).

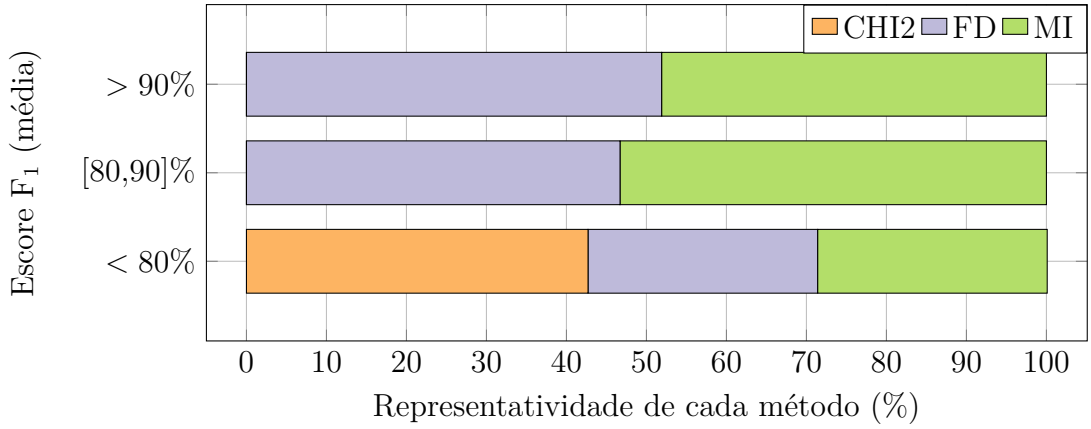


Figura 42: Escore F_1 vs métodos de seleção de características (Unifei 2018).

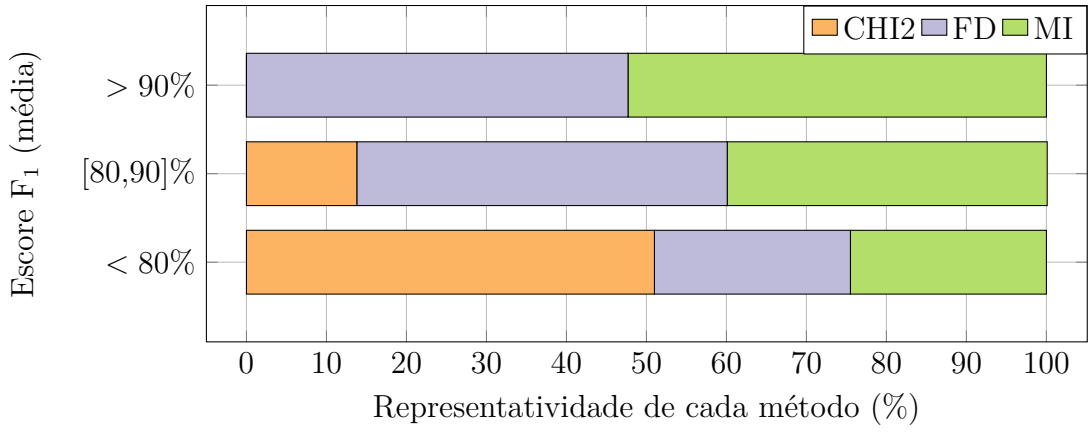


Figura 43: Escore F_1 vs métodos de seleção de características em todas as bases.

Diante dos resultados obtidos, podem ser tiradas as seguintes conclusões:

- Em todos os conjuntos de dados, o método CHI2 foi o que apresentou maior quantidade de padrões zerados. Exceto na base Spam Assassin, esta quantia chegou a representar quase a totalidade (mais de 99%) das mensagens;
- Em todos os conjuntos, o método FD foi o que apresentou menos padrões zerados, tendo o maior percentual (3,6%) ocorrido na base TREC. No caso do MI, o maior percentual (15%) ocorreu nas bases Unifei 2017 e 2018;
- Para todos os métodos de seleção de características e em todos os conjuntos de dados, a quantidade de padrões zerados mostrou-se inversamente proporcional ao número de características das amostras;
- A maioria (51%) dos resultados ruins de escore F_1 médio foi obtida com o método CHI2 (especialmente em Ling Spam e Spam Assassin), seguidos por FD (24,5%) e MI (24,5%);
- A maioria (46,3%) dos resultados médios de escore F_1 médio foi obtida com o método FD (especialmente em TREC e Unifei 2017), seguido por MI (40%) e CHI2 (13,7%);
- A maioria (52,3%) dos resultados bons de escore F_1 médio foi obtida com o método MI (especialmente em Ling Spam, Spam Assassin e TREC), seguido por FD (47,7%). Não houve ocorrências com o método CHI2;
- Nas bases Ling Spam e Spam Assassin, o método MI apresentou os melhores resultados de escore F_1 médio na maioria (54,2% e 57,1%, respectivamente) dos casos;
- Na base TREC, o método MI obteve os melhores resultados de escore F_1 médio em 100% dos casos;
- Nas bases Unifei 2017 e 2018, o método FD apresentou os melhores resultados de escore F_1 médio na maioria (72,2% e 51,9%, respectivamente) dos casos.

5.2 Redução de dimensionalidade

As quantidades de características — NF e NF', antes e após a redução de dimensionalidade, respectivamente — em cada base de *e-mails*, para cada método de seleção, podem ser encontradas nas tabelas 7, 8, 9, 10 e 11 e nas figuras 44, 45, 46, 47 e 48.

Tabela 7: Redução de dimensionalidade na base Ling Spam.

NF	$NF'(CHI2)$	$NF'(FD)$	$NF'(MI)$
8	1 (12,5%)	4 (50%)	6 (75%)
16	1 (6,3%)	4 (25%)	12 (75%)
32	1 (3,1%)	10 (31,3%)	19 (59,4%)
64	3 (4,7%)	19 (29,7%)	34 (53,1%)
128	4 (3,1%)	38 (29,7%)	52 (40,6%)
256	14 (5,5%)	58 (22,7%)	68 (26,6%)
512	14 (2,7%)	103 (20,1%)	139 (27,1%)
1024	242 (23,6%)	121 (11,8%)	157 (15,3%)

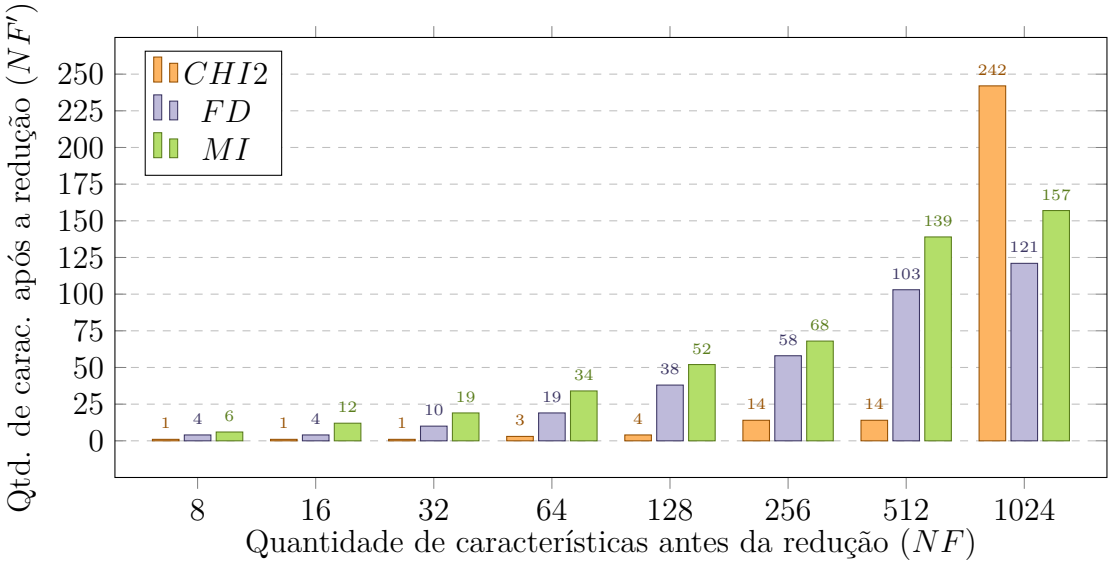


Figura 44: Redução de dimensionalidade na base Ling Spam.

Tabela 8: Redução de dimensionalidade na base Spam Assassin.

NF	$NF'(CHI2)$	$NF'(FD)$	$NF'(MI)$
8	4 (50%)	5 (62,5%)	3 (37,5%)
16	7 (43,8%)	6 (37,5%)	6 (37,5%)
32	14 (43,8%)	9 (28,1%)	18 (56,3%)
64	20 (31,3%)	16 (25%)	33 (51,6%)
128	25 (19,5%)	26 (20,3%)	58 (45,3%)
256	29 (11,3%)	53 (20,7%)	85 (33,2%)
512	7 (1,4%)	67 (13,1%)	96 (18,8%)
1024	20 (2%)	58 (5,7%)	84 (8,2%)

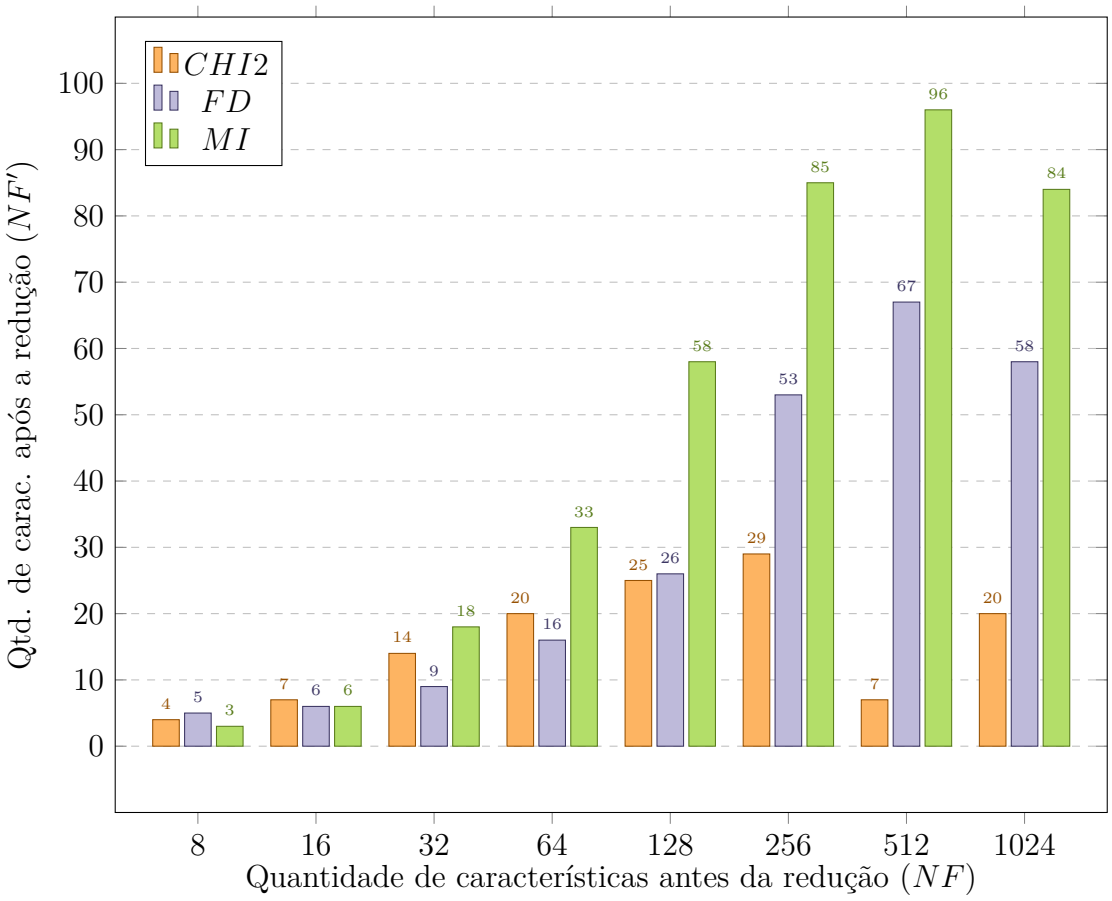


Figura 45: Redução de dimensionalidade na base Spam Assassin.

Tabela 9: Redução de dimensionalidade na base TREC.

NF	$NF'(CHI2)$	$NF'(FD)$	$NF'(MI)$
8	4 (50%)	6 (75%)	5 (62,5%)
16	11 (68,8%)	7 (43,8%)	6 (37,5%)
32	10 (31,3%)	14 (43,8%)	9 (28,1%)
64	12 (18,8%)	7 (10,9%)	16 (25%)
128	45 (35,2%)	22 (17,2%)	26 (20,3%)
256	68 (26,6%)	24 (9,4%)	37 (14,5%)
512	133 (26%)	51 (10%)	66 (12,9%)
1024	47 (4,6%)	49 (4,8%)	204 (19,9%)

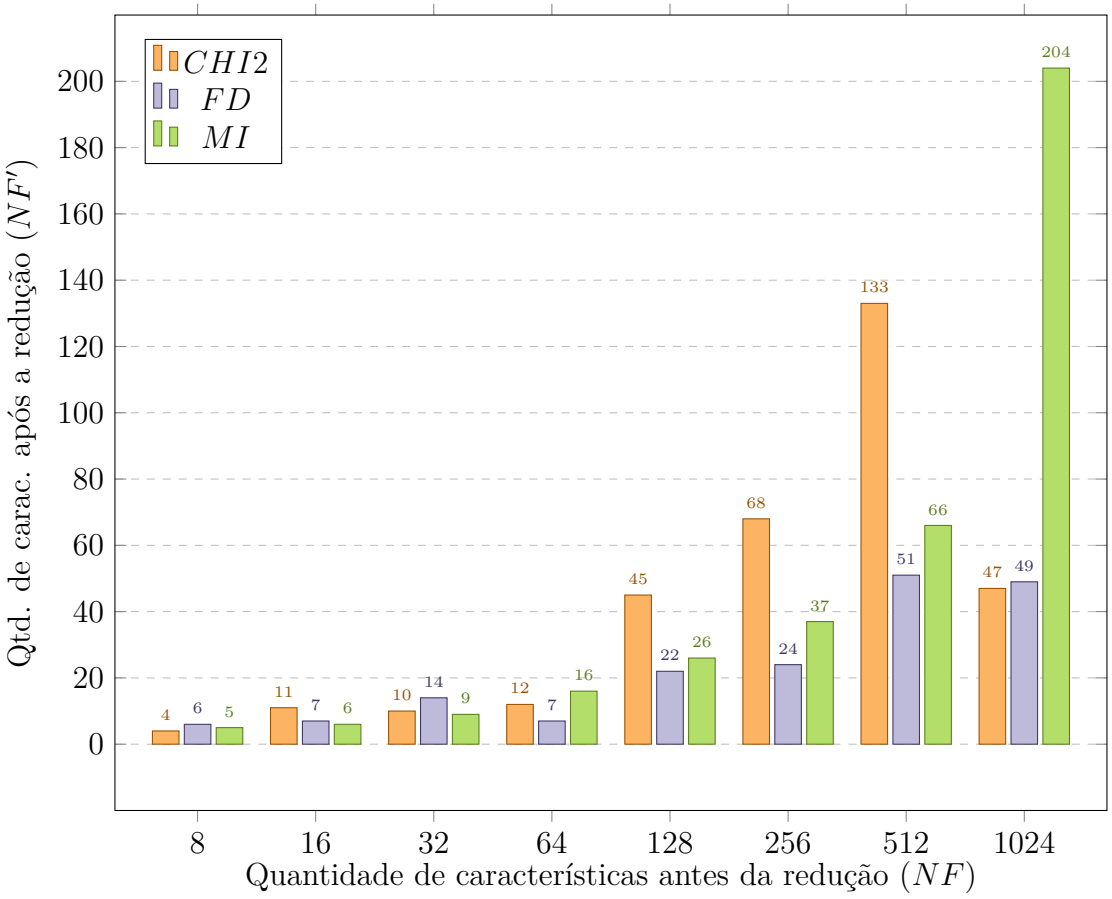


Figura 46: Redução de dimensionalidade na base TREC.

Tabela 10: Redução de dimensionalidade na base Unifei 2017.

NF	$NF'(CHI2)$	$NF'(FD)$	$NF'(MI)$
8	8 (100,0%)	5 (62,5%)	4 (50,0%)
16	10 (62,5%)	3 (18,8%)	7 (43,8%)
32	20 (62,5%)	6 (18,8%)	5 (15,6%)
64	48 (75,0%)	22 (34,4%)	9 (14,1%)
128	86 (67,2%)	65 (50,8%)	18 (14,1%)
256	34 (13,3%)	51 (19,9%)	31 (12,1%)
512	45 (8,8%)	101 (19,7%)	125 (24,4%)
1024	17 (1,7%)	105 (10,3%)	208 (20,3%)

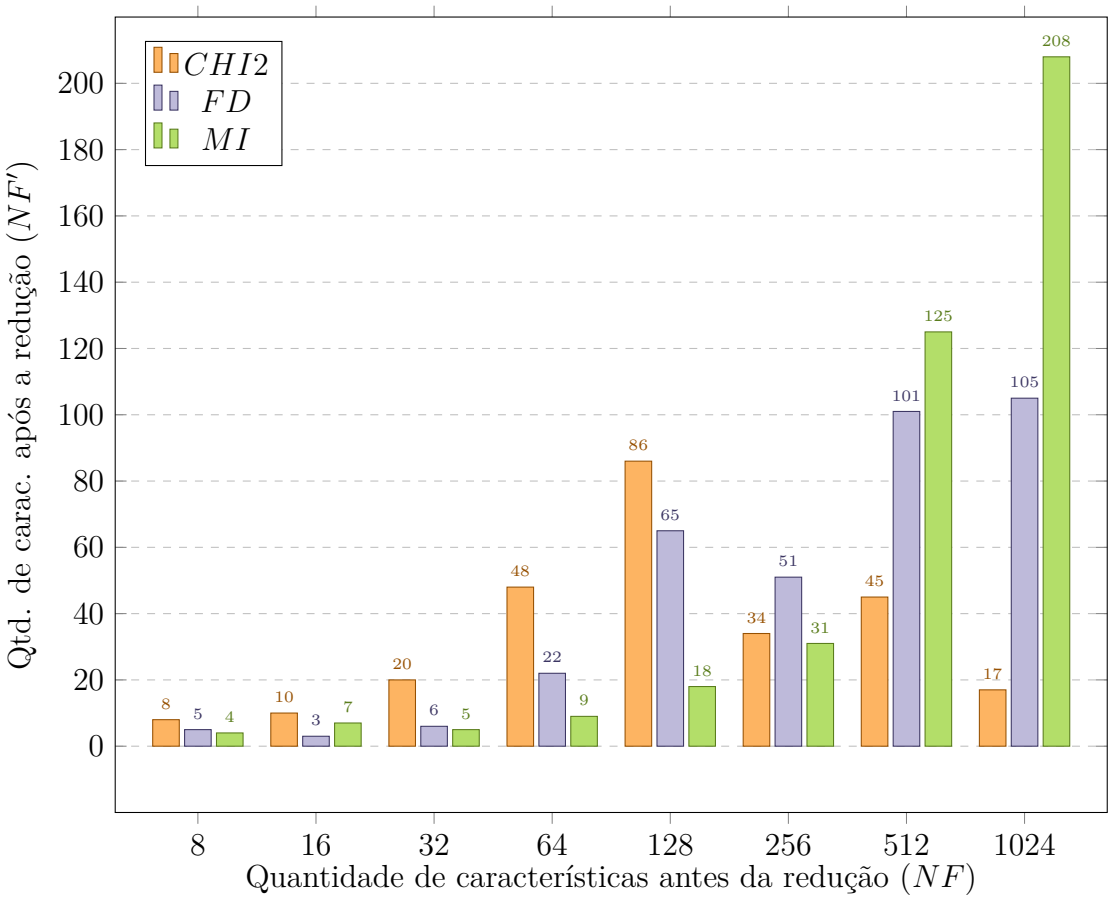


Figura 47: Redução de dimensionalidade na base Unifei 2017.

Tabela 11: Redução de dimensionalidade na base Unifei 2018.

NF	$NF'(CHI2)$	$NF'(FD)$	$NF'(MI)$
8	4 (50%)	5 (62,5%)	3 (37,5%)
16	16 (100%)	3 (18,8%)	3 (18,8%)
32	28 (87,5%)	8 (25%)	7 (21,9%)
64	42 (65,6%)	11 (17,2%)	10 (15,6%)
128	69 (53,9%)	24 (18,8%)	30 (23,4%)
256	116 (45,3%)	60 (23,4%)	49 (19,1%)
512	121 (23,6%)	79 (15,4%)	118 (23%)
1024	152 (14,8%)	77 (7,5%)	155 (15,1%)

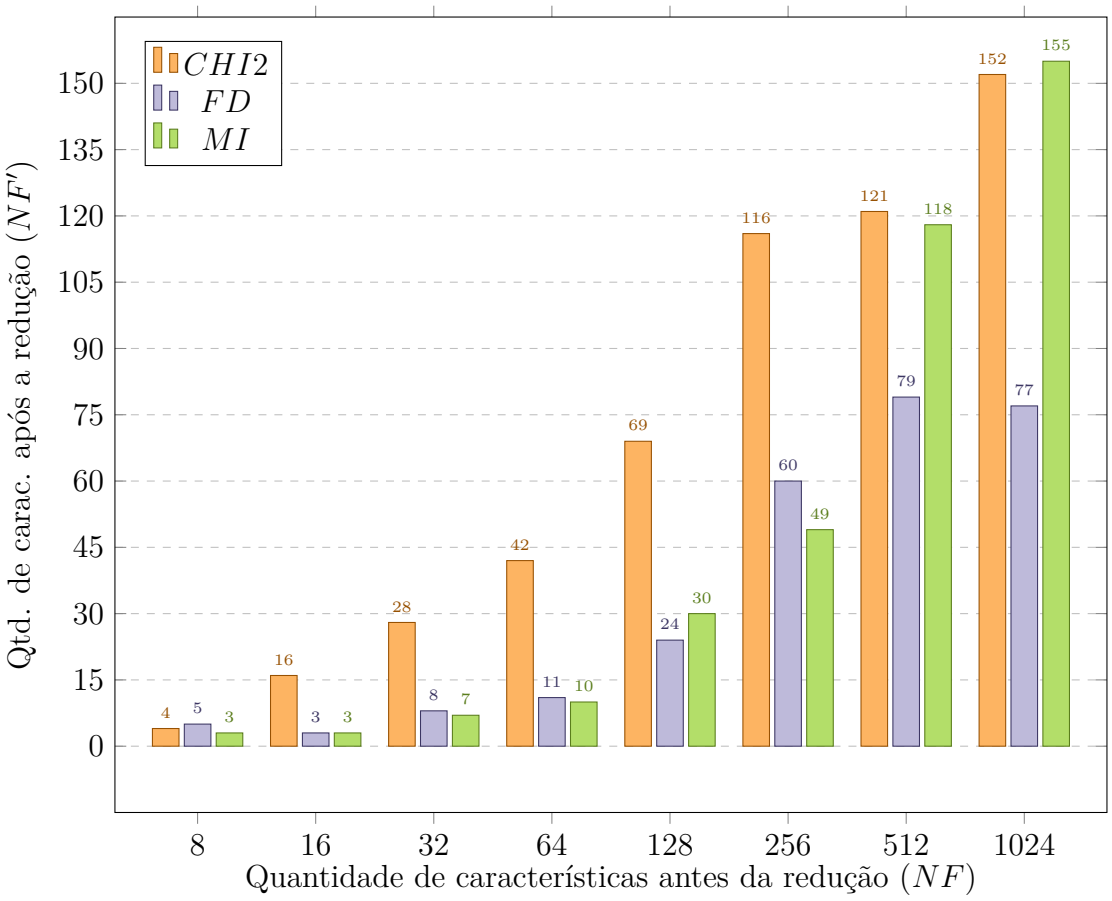


Figura 48: Redução de dimensionalidade na base Unifei 2018.

Diante dos resultados obtidos, podem ser tiradas as seguintes conclusões:

- Nas bases Ling Spam e Spam Assassin, o método CHI2 foi o que mais teve características descartadas pelo algoritmo de redução de dimensionalidade;
- Nas bases TREC, Unifei 2017 e Unifei 2018, o método FD foi o que mais teve características descartadas pelo algoritmo de redução de dimensionalidade;
- Na maioria dos casos, conforme a quantidade de características antes da redução aumentava, a quantidade após a redução também aumentava. Entretanto, algumas anomalias foram detectadas (especialmente quando a dimensionalidade original era de 1024 características):
  - CHI2: bases Spam Assassin (512/7 e 1024/20), TREC (32/10 e 1024/47) e Unifei 2017 (256/34, 512/45 e 1024/17);
  - FD: bases Spam Assassin (1024/58), TREC (64/7 e 1024/49), Unifei 2017 (256/51) e Unifei 2018 (16/3 e 1024/77);
  - MI: bases Spam Assassin (1024/84) e Unifei 2017 (32/5).
- A maior redução percentual para o método CHI2 foi de 98,6% — de 512 (100%) para 7 (1,4%) características —, na base Spam Assassin e as menores foram nas bases Unifei 2017 e 2018, nas quais não houve redução quando a dimensionalidade original era de 8 e 16 características;
- A maior redução percentual para o método FD foi de 95,2% — de 1024 (100%) para 49 (4,8%) características — e a menor foi de 25% — 8 (100%) para 6 (75%) características —, ambas na base TREC;
- A maior redução percentual para o método MI foi de 91,8 % — de 1024 (100%) para 84 (8,2%) características —, na base Spam Assassin e a menor foi de 25 % — de 8 e 16 (100%) para 6 e 12 (75%) características, respectivamente —, na base Ling Spam.

5.3 Treinamento e classificação

Conforme descrito no capítulo 4, cada modelo de aprendizado de máquina é treinado e avaliado dez vezes com cada conjunto de dados e, em seguida, são calculadas as métricas, e seus valores são representados em termos de média e intervalo de confiança. Assim, um experimento foi realizado para cada método e, como já dito, os resultados individuais completos estão disponível nos apêndices.

Os melhores resultados, ordenados por escore F_1 médio, para cada base de *e-mails*, são exibidos nas tabelas 12 a 16 e nas figuras 49 a 53. Nas tabelas, FS é o método de seleção de características e NF e NF' são, respectivamente, as quantidades de características antes e após a redução de dimensionalidade.

Além do *ranking* dos modelos de aprendizado de máquina apresentados nas figuras já mencionadas, resultados complementares podem ser encontrados nas figuras 54 e 55.

O gráfico da figura 54 mostra o quanto cada método de seleção de características esteve presente nos melhores resultados, considerando todos os conjuntos de dados. Esta visualização foi obtida contabilizando quantas vezes cada método (CHI2, FD e MI) apareceu nos *rankings* das tabelas 12, 13, 14, 15 e 16.

De forma análoga, o gráfico da figura 55 mostra o quanto cada quantidade de características antes da redução esteve presente nos melhores resultados, considerando todos os conjuntos de dados. Esta visualização foi obtida contabilizando quantas vezes cada quantidade $(8, 16, \ldots, 1024)$ apareceu nos *rankings* das tabelas 12, 13, 14, 15 e 16.

Tabela 12: *Ranking* dos classificadores na base Ling Spam.

FS	NF	NF'	Método	Escore F_1	Tempo de treinamento	Tempo de classificação
FD	1024	121	AB-M1	99,23 ± 0,07	00:00:05.321 ± 00:00:00.249	00:00:00.018 ± 00:00:00.001
FD	512	103	AODE	98,55 ± 0,05	00:00:00.236 ± 00:00:00.000	00:00:00.418 ± 00:00:00.000
MI	512	139	SLP	98,42 ± 0,21	00:00:06.196 ± 00:00:00.017	00:00:00.011 ± 00:00:00.000
FD	1024	121	REPT	98,26 ± 0,14	00:00:00.308 ± 00:00:00.010	00:00:00.004 ± 00:00:00.000
FD	1024	121	P-SVM	98,20 ± 0,15	00:00:03.431 ± 00:00:00.352	00:00:00.388 ± 00:00:00.002
FD	1024	121	L-SVM	98,16 ± 0,08	00:00:00.250 ± 00:00:00.005	00:00:00.145 ± 00:00:00.002
MI	512	139	RBF	97,01 ± 0,32	00:00:18.229 ± 00:00:00.918	00:00:05.242 ± 00:00:00.021
MI	64	34	NB	95,20 ± 0,50	00:00:00.022 ± 00:00:00.000	00:00:00.098 ± 00:00:00.001

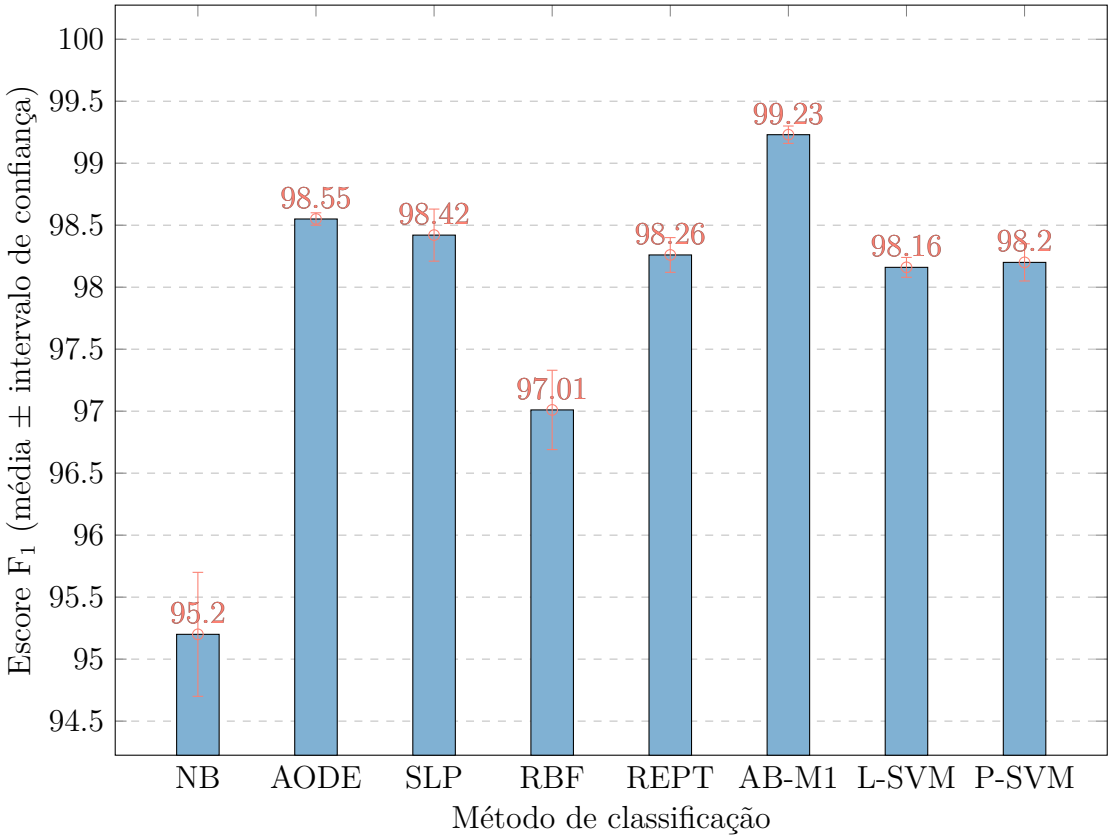


Figura 49: *Ranking* dos classificadores na base Ling Spam.

Tabela 13: *Ranking* dos classificadores na base Spam Assassin.

FS	NF	NF'	Método	Escore F_1	Tempo de treinamento	Tempo de classificação
MI	128	58	AB-M1	96,97 ± 0,29	00:00:00.955 ± 00:00:00.034	00:00:00.007 ± 00:00:00.000
MI	256	85	P-SVM	95,48 ± 0,27	00:00:01.835 ± 00:00:00.091	00:00:00.135 ± 00:00:00.001
MI	128	58	SLP	95,22 ± 0,20	00:00:01.104 ± 00:00:00.029	00:00:00.002 ± 00:00:00.000
MI	128	58	L-SVM	95,04 ± 0,28	00:00:00.087 ± 00:00:00.001	00:00:00.071 ± 00:00:00.003
MI	128	58	AODE	94,70 ± 0,22	00:00:00.037 ± 00:00:00.000	00:00:00.066 ± 00:00:00.000
MI	128	58	REPT	94,04 ± 0,35	00:00:00.064 ± 00:00:00.003	00:00:00.001 ± 00:00:00.000
FD	256	53	RBF	93,38 ± 0,47	00:00:02.782 ± 00:00:00.221	00:00:00.729 ± 00:00:00.002
FD	512	67	NB	89,74 ± 0,37	00:00:00.017 ± 00:00:00.000	00:00:00.078 ± 00:00:00.002

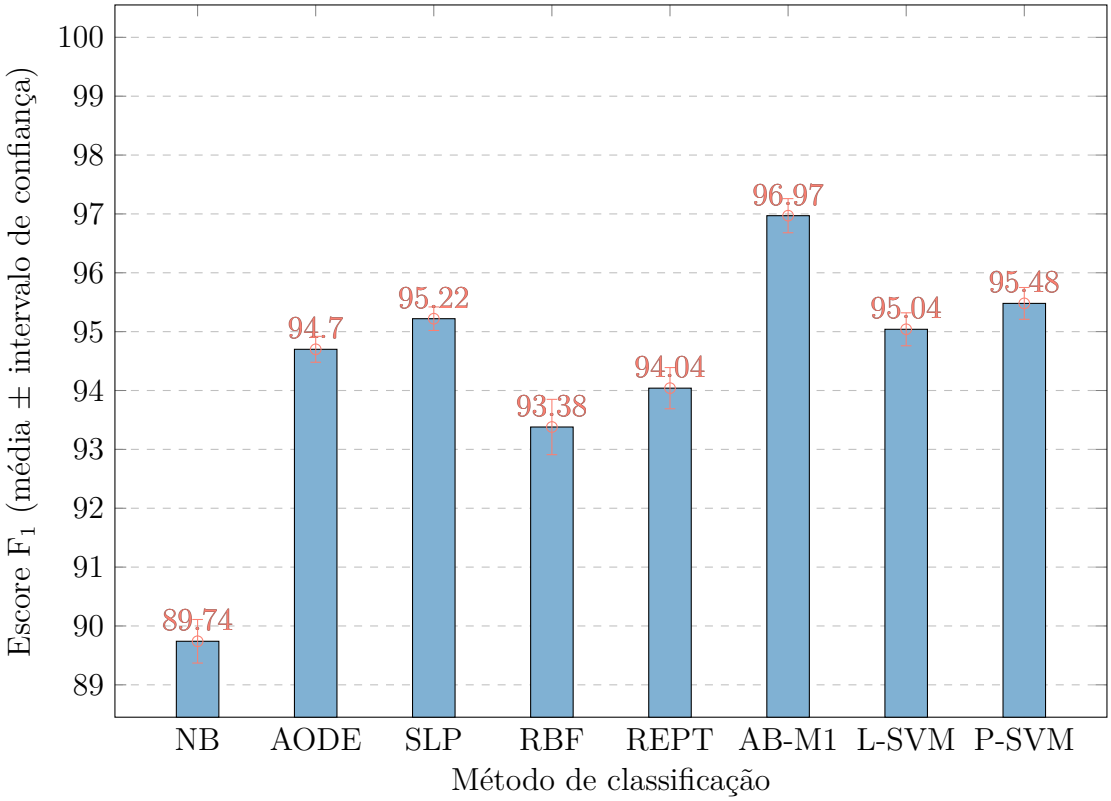


Figura 50: *Ranking* dos classificadores na base Spam Assassin.

Tabela 14: *Ranking* dos classificadores na base TREC.

FS	NF	NF'	Método	Escore F_1	Tempo de treinamento	Tempo de classificação
MI	1024	204	AB-M1	90,63 ± 0,07	00:19:21.268 ± 00:00:29.870	00:00:01.610 ± 00:00:00.104
MI	1024	204	REPT	89,05 ± 0,06	00:01:30.832 ± 00:00:01.972	00:00:00.136 ± 00:00:00.003
MI	1024	204	AODE	87,67 ± 0,07	00:00:13.285 ± 00:00:00.050	00:00:30.387 ± 00:00:00.568
MI	1024	204	P-SVM	86,80 ± 0,11	00:08:49.209 ± 00:01:02.402	00:00:09.555 ± 00:00:00.008
MI	1024	204	L-SVM	82,75 ± 2,73	00:00:03.994 ± 00:00:00.113	00:00:01.943 ± 00:00:00.013
MI	1024	204	SLP	81,67 ± 3,73	00:01:48.535 ± 00:00:00.208	00:00:00.201 ± 00:00:00.000
MI	128	26	RBF	77,20 ± 1,42	00:01:54.845 ± 00:00:24.283	00:00:08.976 ± 00:00:00.026
FD	256	24	NB	72,08 ± 0,70	00:00:00.312 ± 00:00:00.000	00:00:00.813 ± 00:00:00.008

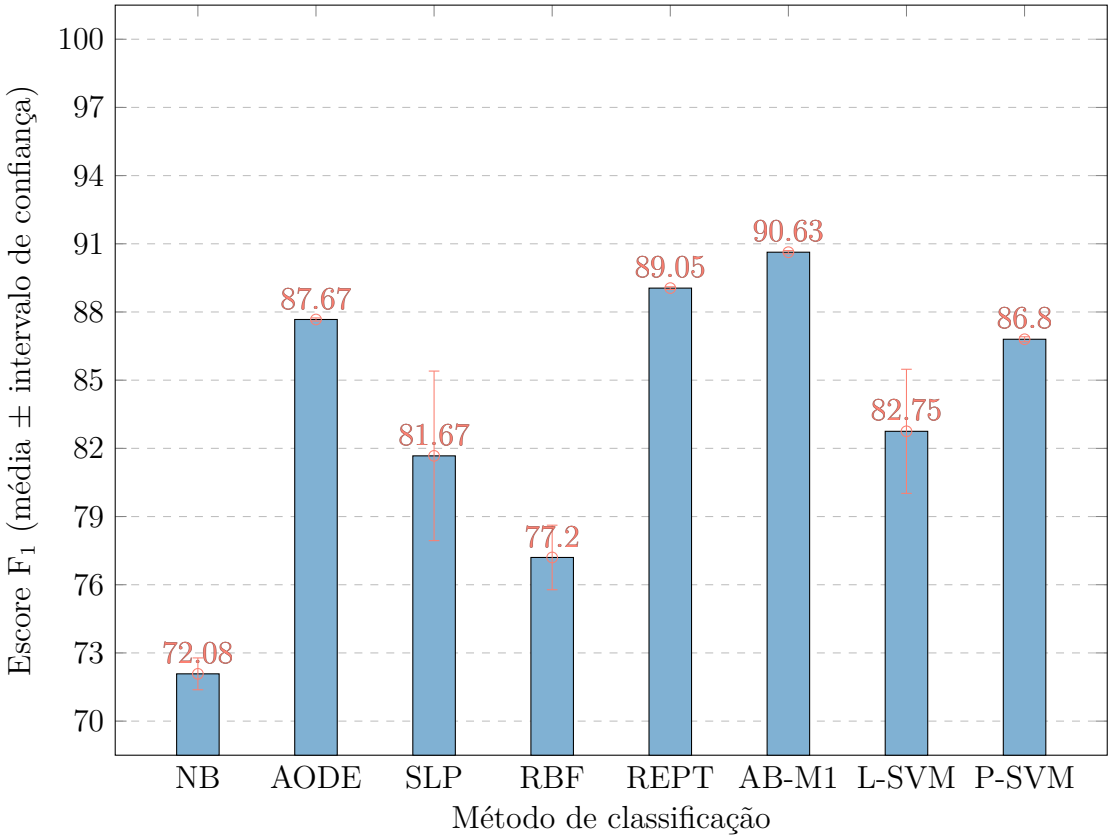


Figura 51: *Ranking* dos classificadores na base TREC.

Tabela 15: *Ranking* dos classificadores na base Unifei 2017.

FS	NF	NF'	Método	Escore F_1	Tempo de treinamento	Tempo de classificação
FD	128	65	AB-M1	93,96 ± 0,02	00:17:11.884 ± 00:00:38.035	00:00:03.276 ± 00:00:00.121
FD	128	65	REPT	93,27 ± 0,02	00:01:18.676 ± 00:00:01.901	00:00:00.356 ± 00:00:00.002
FD	128	65	AODE	92,37 ± 0,06	00:00:29.522 ± 00:00:00.336	00:00:17.162 ± 00:00:00.255
MI	1024	208	P-SVM	83,00 ± 0,09	01:32:39.551 ± 00:28:47.995	00:00:40.832 ± 00:00:00.019
MI	1024	208	L-SVM	80,91 ± 0,41	00:00:22.785 ± 00:00:01.653	00:00:08.090 ± 00:00:00.028
MI	1024	208	SLP	78,78 ± 1,32	00:07:43.886 ± 00:00:00.809	00:00:00.839 ± 00:00:00.011
MI	512	125	RBF	77,52 ± 0,23	00:24:12.225 ± 00:03:07.822	00:04:07.769 ± 00:00:00.561
MI	512	125	NB	75,44 ± 0,17	00:00:07.338 ± 00:00:00.064	00:00:16.914 ± 00:00:00.034

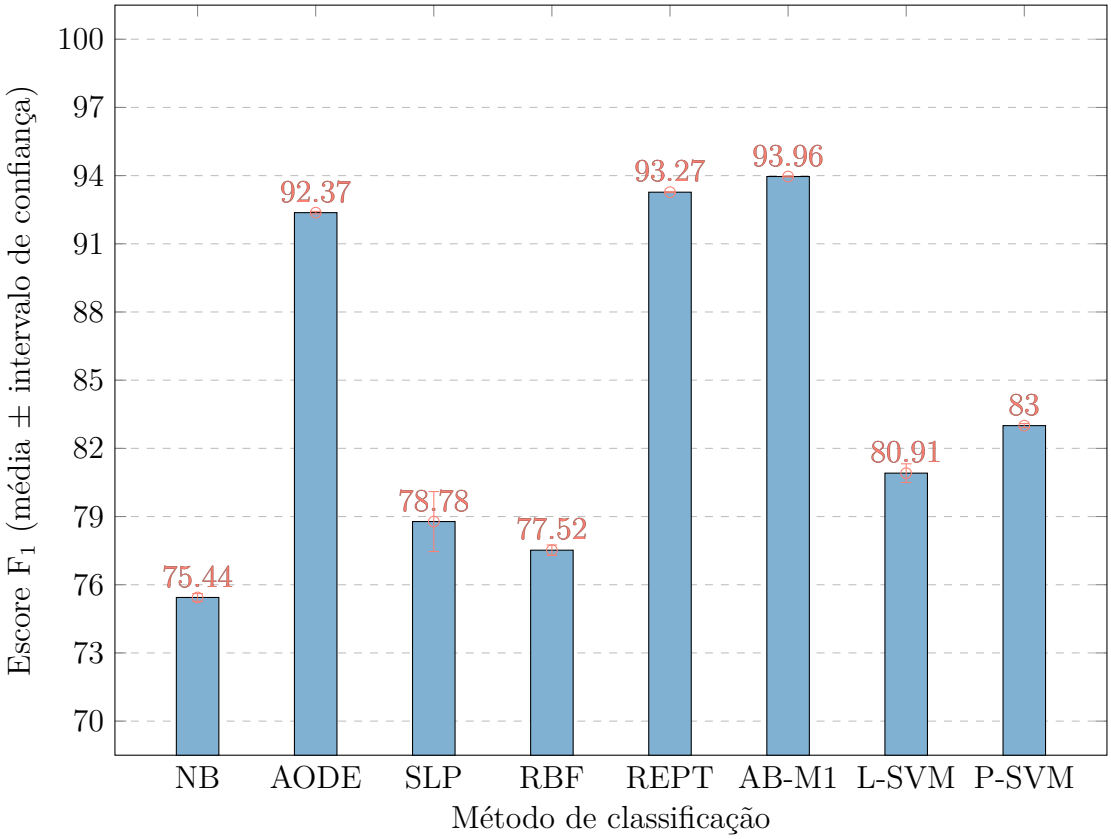


Figura 52: *Ranking* dos classificadores na base Unifei 2017.

Tabela 16: *Ranking* dos classificadores na base Unifei 2018.

FS	NF	NF'	Método	Escore F_1	Tempo de treinamento	Tempo de classificação
FD	8	5	AB-M1	97,80 ± 0,03	00:00:39.308 ± 00:00:01.229	00:00:02.491 ± 00:00:00.066
FD	8	5	REPT	96,66 ± 0,02	00:00:04.153 ± 00:00:00.014	00:00:00.293 ± 00:00:00.003
FD	8	5	AODE	95,07 ± 0,04	00:00:02.406 ± 00:00:00.028	00:00:00.747 ± 00:00:00.040
MI	1024	155	P-SVM	82,27 ± 0,17	00:58:43.293 ± 00:12:58.240	00:00:25.196 ± 00:00:00.013
MI	1024	155	L-SVM	81,23 ± 0,30	00:00:14.037 ± 00:00:00.911	00:00:06.054 ± 00:00:00.018
MI	32	7	RBF	79,24 ± 1,04	00:01:46.418 ± 00:00:03.602	00:00:07.948 ± 00:00:00.026
MI	512	118	SLP	79,17 ± 4,57	00:04:52.090 ± 00:00:00.531	00:00:00.540 ± 00:00:00.012
FD	512	79	NB	71,96 ± 1,17	00:00:04.986 ± 00:00:00.061	00:00:10.410 ± 00:00:00.082

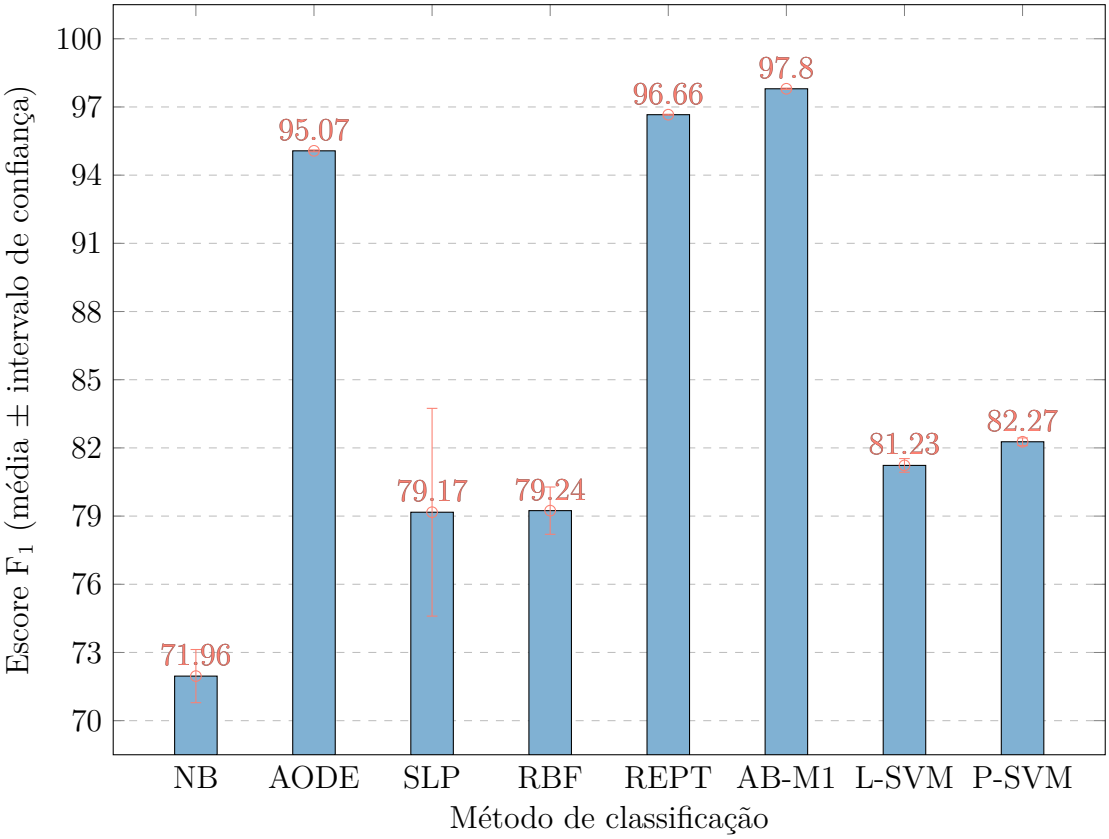


Figura 53: *Ranking* dos classificadores na base Unifei 2018.

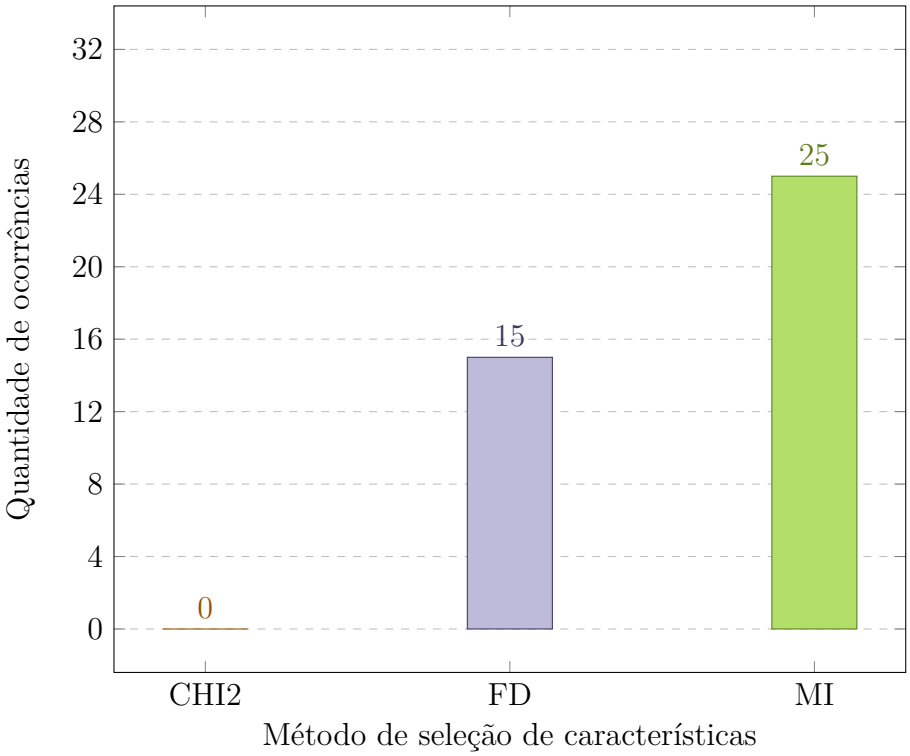


Figura 54: Distribuição dos métodos de seleção de características no *ranking*.

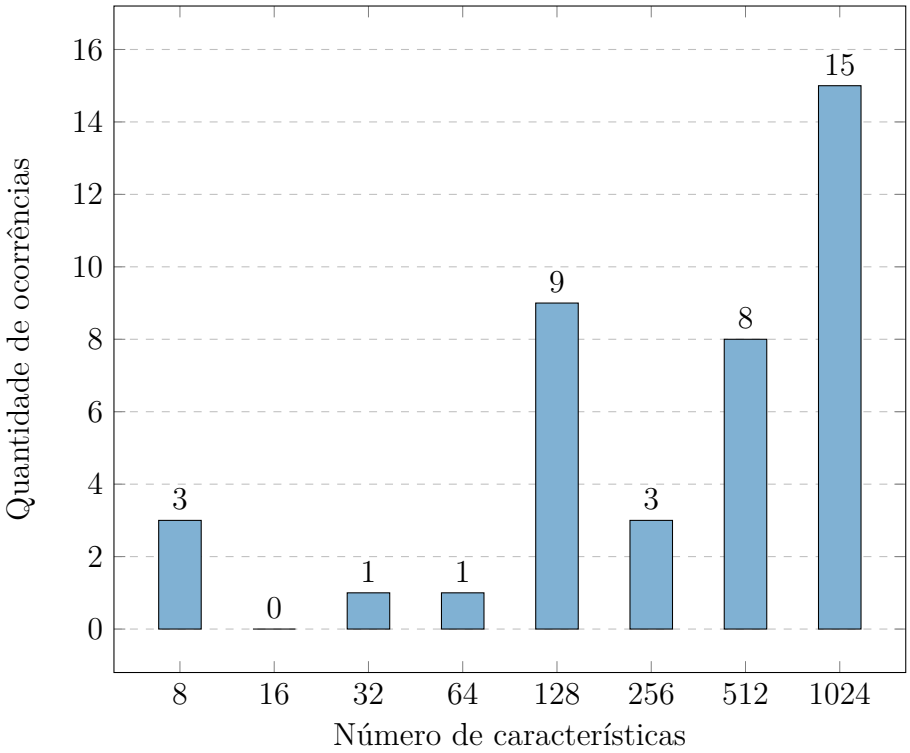


Figura 55: Distribuição das quantidades de características no *ranking*.

Diante destes resultados e levando-se em conta empates técnicos devido à sobreposição de intervalos de confiança, podem ser tiradas as seguintes conclusões:

- Escore F_1
  - O classificador baseado em árvore de decisão e *boosting* (AB-M1) ocupou o primeiro lugar em todos os conjuntos de dados. Portanto, superou a abordagem sem *boosting* (REPT) em todos os cenários, que ocupou os quarto e quinto lugares nas bases menores (Ling Spam e Spam Assassin), e terceiro lugar nas demais (TREC, Unifei 2017 e 2018).
  - O classificador probabilístico não-tão-ingênuo (AODE) superou o ingênuo (NB) em todos os conjuntos de dados. Nas bases menores (Ling Spam, Spam Assassin) e na média (TREC), o AODE ficou em segundo e quinto lugar, respectivamente, e nas demais (Unifei 2017 e 2018), ocupou sempre o terceiro lugar. Em contrapartida, o NB ficou em último colocado em todos os conjuntos de dados.
  - O classificador neural de camada única (SLP) superou a rede de função de base radial (RBF) em todas as bases com exceção da Unifei 2018, ocupando o terceiro lugar nas bases Ling Spam e Spam Assassin, sexto lugar nas bases TREC e Unifei 2017 e sétimo lugar na Unifei 2018. Em contrapartida, o RBF ocupou o sétimo lugar em todas as bases, com exceção da Unifei 2018, onde ocupou a sexta posição.
  - Em todos os conjuntos de dados, a máquina de vetores de suporte de *kernel* polinomial (P-SVM) apresentou resultado superior à de *kernel* linear (L-SVM). A versão polinomial ficou em quinto e segundo lugar nas bases Ling Spam e Spam Assassin, respectivamente, e em quarto lugar nas bases TREC, Unifei 2017 e 2018. Já a versão linear ocupou a sexta posição na base Ling Spam, quarta posição na Spam Assassin e quinto lugar nas demais.

- Tempo de treinamento
  - Nos conjuntos de dados menores (Ling Spam e Spam Assassin), o método RBF apresentou o maior tempo de treinamento. Na base média (TREC), essa posição foi ocupada pela árvore com *boosting* (AB-M1) e, nas bases maiores (Unifei 2017 e 2018), pela SVM de núcleo polinomial.
  - O método NB apresentou o menor tempo de treinamento em todos os conjuntos, com exceção do Unifei 2018, onde o treinamento mais rápido foi da sua contraparte não-tão-ingênua (AODE).
- Tempo de classificação
  - O método RBF apresentou o maior tempo de classificação nos conjuntos Ling Spam, Spam Assassin e Unifei 2017. Nas bases TREC e Unifei 2018, os métodos AODE e P-SVM foram os mais demorados, respectivamente.
  - Em todos os conjuntos de dados, a árvore de decisão sem *boosting* (REPT) apresentou o menor tempo de classificação.
- Seleção de características e redução de dimensionalidade
  - O método de seleção de características MI foi o que teve maior percentual de ocorrências no *ranking* (62,5%), seguido pelo FD (37,5%). Não houve registros nesse sentido para método CHI2.
  - A dimensionalidade original de 1024 características foi a que teve maior percentual de ocorrências no *ranking* (37,5%), seguida por 128 (22,5%) e 512 (20%). As demais quantidades somam 20% das ocorrências, e não houve ocorrências para a configuração com 16 características no *ranking*.

6 Conclusão

6.1 Discussão dos resultados e considerações finais

Nos dias de hoje, com o advento das redes de dados móveis, um número cada vez maior de pessoas possui acesso à Internet de qualquer lugar. Seja no trabalho, em casa ou até mesmo no trânsito, é possível receber e enviar *e-mails* a qualquer momento. Isto implica em um aumento na quantidade de usuários de *e-mail* e, consequentemente, na quantidade de trocas de mensagens. Infelizmente, não são somente a quantidade de usuários e a facilidade de acesso que têm aumentado. Desde 2005, pelo menos metade do tráfego de *e-mails* é composto por *spam*.

Qualquer *e-mail* não-solicitado é considerado *spam*. Geralmente, este tipo de mensagem é enviada a um grande número de pessoas para promover um produto ou serviço. Isso inclui anúncios, esquemas de pirâmide, doações, correntes de mensagens, *e-mail* político, assessoria no mercado de ações e avisos únicos.

Este tráfego massivo de mensagens indesejadas provoca consequências negativas, tais como o uso excessivo da largura de banda, impacto ambiental, perda de produtividade, exposição de conteúdo impróprio a públicos sensíveis, prejuízos financeiros e legais e redução da efetividade de propagandas legítimas.

Uma técnica largamente empregada por servidores de *e-mail*, a fim de lidar com o problema de *spam*, é através de *listas negras*. Uma lista negra de *e-mail* é um banco de dados em tempo real que determina se um endereço IP está enviando *e-mails* que podem ser considerados *spam*. Este tipo de sistema apresenta diversas

desvantagens, tais como código fechado, parcialidade questionável, limitações devido aos limites dos protocolos IPv4/IPv6 e lentidão na classificação.

Uma outra abordagem ao problema de classificação de *e-mails* consiste em usar os dados contidos na própria mensagem (isto é, remetente, destinatário, assunto, cabeçalho, corpo, anexos etc.) e um modelo de aprendizado de máquina a fim de construir um sistema de decisão automático, inteligente e imparcial. Estes devem ser capazes de identificar padrões e de se ajustarem à características comuns a ambos os tipos de *e-mail*, resultando em classificações cada vez mais precisas, com o decorrer das épocas de treinamento.

Portanto, o objetivo geral deste trabalho é aplicar modelos de aprendizado de máquina à classificação de *e-mails*, a fim de obter modelos utilizáveis em sistemas anti-*spam* do mundo real. Para isto, foi necessário o estudo de métodos de seleção de características e de redução de dimensionalidade, bem como uma revisão teórica de diversos algoritmos usados na categorização de texto. Em seguida, foi elaborado um *framework* de *software* científico genérico, capaz de treinar diversos modelos de aprendizado de máquina e salvá-los para posterior reuso.

Após extensa simulação sobre três conjuntos de dados públicos (Ling Spam, Spam Assassin e TREC) e dois privados (Unifei 2017 e Unifei 2018), os resultados experimentais de cada combinação de modelo de aprendizado de máquina e método de seleção de características foram coletados, consolidados e avaliados em termos de métricas bem difundidas na comunidade científica, tais como precisão, revocação e escore F_1, dados em termos de média e intervalo de confiança entre as repetições.

Em termos de seleção de características e padrões zerados, constatou-se que os métodos CHI2 e FD foram os que apresentaram maior e menor quantidade de padrões zerados, respectivamente. Além disso, em todos os conjuntos de dados e para todos os métodos de seleção de características, observou-se que a quantidade de padrões zerados foi inversamente proporcional ao número de características das amostras. Por fim, concluiu-se que a grande maioria dos resultados ruins foi obtida com o método CHI2 e a maioria dos resultados bons foi obtida com o método MI.

Em termos de redução de dimensionalidade, os métodos CHI2 e FD foram o que tiveram as maiores quantidades de características descartadas nas bases menores (Ling Spam e Spam Assassin) e maiores (TREC, Unifei 2017 e Unifei 2018), respectivamente. Além disso, observou-se que a redução de até 512 características foi crescente em praticamente todos os casos, sofrendo anomalias especialmente quando a dimensionalidade original era de 1024 características.

Em termos de desempenho na classificação, um método baseado em árvores de decisão "impulsionadas" (AB-M1) mostrou excelentes taxas de acerto em todos os conjuntos de dados. Além disso, foi mostrado que um classificador baseado em redes neurais simples (SLP) foi capaz que apresentar resultados superiores a um modelo mais elaborado (RBF). Neste sentido, também foi possível observar que a suposição de independência em métodos probabilísticos (NB) resultou em desempenho inferior à suposição de dependência restrita (AODE). Finalmente, notou-se que a máquina de vetores de suporte de núcleo polinomial (P-SVM) apresentou resultado superior à de núcleo linear (L-SVM) em todos os casos.

Em termos dos tempos de processamento, enquanto os métodos probabilísticos (ingênuo — NB — e de dependência restrita — AODE) foram treinados mais rapidamente em todas as bases, o modelo de árvore de decisão "impulsionada" (AB-M1) e a máquina de vetores de núcleo polinomial apresentaram os maiores tempos de treinamento. Para classificar as amostras, entretanto, o método que se mostrou mais rápido em todos os casos foi a árvore de decisão simples (REPT). Finalmente, os que mais demoraram para efetuar a classificação variaram, sendo a rede neural de função de base radial (RBF) a mais frequente.

Os resultados experimentais confirmam a hipótese de que um modelo de aprendizado de máquina (especialmente aqueles compostos por árvores de decisão "impulsionadas"), aliado ao método de seleção de características baseado em informação mútua e à técnica de redução de dimensionalidade baseada em correlação pode ser aplicado com sucesso à categorização automática de *e-mails*, protegendo os usuários de conteúdo impróprio e/ou malicioso com alto grau de confiabilidade.

6.2 Sugestões para novos trabalhos

Tanto na etapa de pesquisas e revisões — bibliográficas e de fundamentação teórica — preliminares, quanto durante escrita deste documento, foram identificados tópicos e pontos de possíveis melhora e/ou extensão de pesquisa. Os mais notáveis, identificados como sugestões para novos trabalhos, foram discriminados na listagem a seguir, bem como referências relevantes a cada item.

- Uso de outros métodos de seleção de características. Exemplos: *Gini Index* (GINI, 1912), BNS (do inglês *Bi-Normal Separation*), IG e *Odds Ratio* (FORMAN; GEORGE, 2003).
- Exploração do espaço e otimização de parâmetros de configuração de cada modelo de aprendizado de máquina. (CHAPELLE et al., 2002)
- Uso de técnicas de neuroevolução. Exemplos: GNARL (do inglês *GeNeralized Acquisition of Recurrent Links*) (ANGELINE; SAUNDERS; POLLACK, 1994), NEAT (do inglês *NeuroEvolution of Augmenting Topologies*) (STANLEY; MIIKKULAINEN, 2002) e EANT2 (do inglês *Evolutionary Acquisition of Neural Topologies, Version 2*) (SIEBEL; SOMMER, 2007).
- Uso de bibliotecas de aprendizado de máquina e/ou plataformas computacionais com suporte a paralelismo. Exemplo: Nvidia CUDA (do inglês *Compute Unified Device Architecture*) (KIRK, 2007), que utiliza GPUs (do inglês *Graphics Processing Unit*) dedicadas para acelerar o processamento de problemas com grande capacidade de paralelização.
- Uso de técnicas de *deep learning* (DENG; YU, 2014).

7 Anexos

O código do *framework* desenvolvido[1] e os conjuntos de dados utilizados nos experimentos[2] foram disponibilizados sob a licença GNU GENERAL PUBLIC LICENSE versão 3, de 29 de junho de 2007[3].

```
This program is free software: you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation, either version 3 of the License, or (at
your option) any later version.

This program is distributed in the hope that it will be useful, but
WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTA-
BILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public
License for more details.

You should have received a copy of the GNU General Public License
along with this program. If not, see <http://www.gnu.org/licenses/>.
```

[1] Disponível em <https://github.com/marcelovca90/anti-spam-weka-cli>
[2] Disponível em <https://github.com/marcelovca90/anti-spam-weka-data>
[3] Disponível em <https://www.gnu.org/licenses/gpl-3.0.html>

8 Apêndices

A tabela 17 apresenta os parâmetros de configuração de cada modelo de aprendizado de máquina utilizado no WEKA.

As tabelas 18, 19, 20, 21, 22, 23, 24, e 25 mostram os resultados completos das execuções dos classificadores NB, AODE, SLP, RBF, REPT, AB-M1, L-SVM e P-SVM, respectivamente.

As tabelas 26 e 27 mostram os detalhes de *hardware* (processador, memória e sistema operacional) e de *software* (Java, WEKA e bibliotecas/dependências), respectivamente, utilizados na execução dos experimentos.

Tabela 17: Parâmetros de configuração usados no WEKA.

Método	Parâmetro	Descrição	Valor [a] [b] [c]
NB	N/D	N/D	N/D
AODE	N/D	N/D	N/D
SLP	-I	quantidade de iterações de treinamento	500
	-L	taxa de aprendizado	0.1
	-B	valor do limiar de ativação	1.0
	-R	semente do gerador de números aleatórios	1
	-M	função de taxa de aprendizado	*LD*
RBF	-B	número de agrupamentos a serem gerados	$(\log_2 NF)^2$
	-S	semente do gerador de números aleatórios	1
	-R	valor de pico (*ridge*) para a regressão	10^{-8}
	-M	limite de iterações da regressão logística	-1
	-W	desvio padrão mínimo para os agrupamentos	0.1
REPT	-M	peso total mínimo de amostras em uma folha	2
	-V	menor proporção da variância de todos os dados que deve estar presente em um nó para realizar divisão em árvores de regressão	0.001
	-N	quantidade de frações dos dados de treinamento usadas na poda da árvore de decisão	3
	-S	semente do gerador de números aleatórios	1
	-L	profundidade máxima da árvore	-1
AB-M1	-P	limiar de peso para poda	100
	-S	semente do gerador de números aleatórios	1
	-I	quantidade de iterações de *boosting*	10
	-W	classificador sobre o qual será feito o *boosting*	*REPT*
L-SVM	-s	tipo do algoritmo solucionador	2
	-c	parâmetro de regularização	128.0
P-SVM	-s	tipo do algoritmo solucionador	2
	-c	parâmetro de regularização	128.0
	-g	gamma da função de mapeamento	1
	-r	coef0 da função de mapeamento	1

[a] NF é a quantidade original de características (ou seja, antes da redução de dimensionalidade)
[b] LD é uma função que decrementa a taxa de aprendizado com o decorrer do treinamento
[c] *REPT* é o classificador *Reduced-Error Pruning Tree*, explicado com detalhes em 2.6.3.1

Tabela 18: Resultados do classificador NB.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.000 ± 00:00:00.000	00:00:00.008 ± 00:00:00.001
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	77,42 ± 0,04	00:00:00.000 ± 00:00:00.000	00:00:00.008 ± 00:00:00.000
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	78,40 ± 0,03	00:00:00.000 ± 00:00:00.000	00:00:00.008 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,30 ± 0,14	99,94 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.000 ± 00:00:00.000	00:00:00.021 ± 00:00:00.002
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,47 ± 0,09	99,92 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.001 ± 00:00:00.000	00:00:00.031 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,74 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:00.002 ± 00:00:00.000	00:00:00.083 ± 00:00:00.002
Ling Spam	CHI2	512	14	68,27 ± 0,16	90,68 ± 0,33	97,35 ± 0,10	36,26 ± 0,51	68,45 ± 0,25	00:00:00.004 ± 00:00:00.000	00:00:00.076 ± 00:00:00.002
Ling Spam	CHI2	1024	242	98,64 ± 0,27	44,88 ± 0,16	17,90 ± 0,31	99,63 ± 0,08	42,98 ± 0,25	00:00:00.077 ± 00:00:00.000	00:00:01.269 ± 00:00:00.002
Ling Spam	FD	8	4	96,82 ± 0,38	73,18 ± 0,41	60,73 ± 0,83	98,17 ± 0,23	79,44 ± 0,42	00:00:00.003 ± 00:00:00.000	00:00:00.015 ± 00:00:00.000
Ling Spam	FD	16	4	96,84 ± 0,37	73,31 ± 0,43	60,99 ± 0,86	98,18 ± 0,22	79,59 ± 0,44	00:00:00.003 ± 00:00:00.000	00:00:00.016 ± 00:00:00.000
Ling Spam	FD	32	10	97,46 ± 0,28	81,61 ± 0,35	75,85 ± 0,53	98,19 ± 0,20	87,30 ± 0,31	00:00:00.007 ± 00:00:00.000	00:00:00.032 ± 00:00:00.000
Ling Spam	FD	64	19	97,25 ± 0,20	86,37 ± 0,29	83,14 ± 0,35	97,84 ± 0,16	90,74 ± 0,13	00:00:00.012 ± 00:00:00.000	00:00:00.061 ± 00:00:00.000
Ling Spam	FD	128	38	96,37 ± 0,26	90,71 ± 0,32	89,17 ± 0,37	96,93 ± 0,20	93,20 ± 0,24	00:00:00.025 ± 00:00:00.000	00:00:00.110 ± 00:00:00.002
Ling Spam	FD	256	58	96,72 ± 0,39	91,99 ± 0,26	90,76 ± 0,26	97,18 ± 0,34	94,10 ± 0,25	00:00:00.036 ± 00:00:00.000	00:00:00.168 ± 00:00:00.003
Ling Spam	FD	512	103	95,42 ± 0,27	94,16 ± 0,20	93,50 ± 0,20	95,89 ± 0,22	94,75 ± 0,10	00:00:00.065 ± 00:00:00.000	00:00:00.284 ± 00:00:00.004
Ling Spam	FD	1024	121	93,95 ± 0,45	92,50 ± 0,35	91,63 ± 0,40	94,59 ± 0,41	93,17 ± 0,25	00:00:00.071 ± 00:00:00.000	00:00:00.333 ± 00:00:00.001
Ling Spam	MI	8	6	70,08 ± 0,31	95,56 ± 0,17	96,85 ± 0,13	62,11 ± 0,54	78,17 ± 0,26	00:00:00.004 ± 00:00:00.000	00:00:00.022 ± 00:00:00.000
Ling Spam	MI	16	12	98,47 ± 0,13	84,22 ± 0,19	79,78 ± 0,40	98,86 ± 0,10	89,61 ± 0,17	00:00:00.008 ± 00:00:00.000	00:00:00.037 ± 00:00:00.000
Ling Spam	MI	32	19	86,76 ± 1,07	97,42 ± 0,13	97,50 ± 0,16	86,32 ± 1,35	91,66 ± 0,65	00:00:00.013 ± 00:00:00.000	00:00:00.056 ± 00:00:00.000
Ling Spam	MI	64	34	96,18 ± 1,27	94,56 ± 1,93	93,78 ± 2,39	96,49 ± 1,29	95,20 ± 0,50	00:00:00.022 ± 00:00:00.000	00:00:00.098 ± 00:00:00.001
Ling Spam	MI	128	52	98,00 ± 0,27	91,21 ± 0,21	89,66 ± 0,31	98,32 ± 0,22	94,16 ± 0,22	00:00:00.033 ± 00:00:00.000	00:00:00.143 ± 00:00:00.001
Ling Spam	MI	256	68	98,24 ± 0,19	91,48 ± 0,29	89,98 ± 0,37	98,53 ± 0,16	94,42 ± 0,18	00:00:00.042 ± 00:00:00.000	00:00:00.188 ± 00:00:00.002
Ling Spam	MI	512	139	97,47 ± 0,24	91,99 ± 0,26	90,70 ± 0,31	97,85 ± 0,20	94,41 ± 0,20	00:00:00.086 ± 00:00:00.000	00:00:00.370 ± 00:00:00.001
Ling Spam	MI	1024	157	95,80 ± 0,35	90,72 ± 0,29	89,23 ± 0,38	96,42 ± 0,28	92,97 ± 0,29	00:00:00.116 ± 00:00:00.000	00:00:00.424 ± 00:00:00.004
Spam Assassin	CHI2	8	4	59,42 ± 0,54	78,10 ± 0,44	85,46 ± 0,25	47,07 ± 0,51	64,14 ± 0,35	00:00:00.001 ± 00:00:00.000	00:00:00.005 ± 00:00:00.000
Spam Assassin	CHI2	16	7	61,19 ± 0,47	79,78 ± 0,56	85,77 ± 0,28	50,80 ± 0,59	66,52 ± 0,43	00:00:00.001 ± 00:00:00.000	00:00:00.009 ± 00:00:00.000
Spam Assassin	CHI2	32	14	62,86 ± 0,30	80,44 ± 0,36	85,39 ± 0,29	54,35 ± 0,47	68,45 ± 0,26	00:00:00.003 ± 00:00:00.000	00:00:00.025 ± 00:00:00.004
Spam Assassin	CHI2	64	20	63,07 ± 0,66	81,26 ± 0,26	85,83 ± 0,23	55,01 ± 1,07	68,96 ± 0,58	00:00:00.004 ± 00:00:00.000	00:00:00.031 ± 00:00:00.000
Spam Assassin	CHI2	128	25	66,22 ± 0,44	81,99 ± 0,47	85,32 ± 0,78	60,52 ± 1,06	71,98 ± 0,32	00:00:00.005 ± 00:00:00.000	00:00:00.037 ± 00:00:00.000
Spam Assassin	CHI2	256	29	93,28 ± 0,70	59,12 ± 0,45	25,39 ± 0,68	98,33 ± 0,19	57,66 ± 0,60	00:00:00.006 ± 00:00:00.000	00:00:00.043 ± 00:00:00.002
Spam Assassin	CHI2	512	7	67,03 ± 9,48	76,50 ± 14,74	58,87 ± 34,32	66,01 ± 20,37	55,44 ± 10,42	00:00:00.001 ± 00:00:00.000	00:00:00.012 ± 00:00:00.001
Spam Assassin	CHI2	1024	20	66,05 ± 0,72	81,17 ± 1,25	85,06 ± 1,31	59,48 ± 1,63	71,37 ± 0,67	00:00:00.004 ± 00:00:00.000	00:00:00.036 ± 00:00:00.000
Spam Assassin	FD	8	5	58,29 ± 0,32	94,26 ± 0,96	97,62 ± 0,42	35,85 ± 0,22	62,02 ± 0,21	00:00:00.001 ± 00:00:00.000	00:00:00.008 ± 00:00:00.004
Spam Assassin	FD	16	6	62,43 ± 0,64	90,28 ± 0,83	94,45 ± 0,56	47,53 ± 1,12	68,45 ± 0,68	00:00:00.002 ± 00:00:00.000	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	32	9	62,84 ± 0,68	89,95 ± 0,73	94,11 ± 0,49	48,62 ± 1,37	68,98 ± 0,75	00:00:00.002 ± 00:00:00.000	00:00:00.012 ± 00:00:00.002
Spam Assassin	FD	64	16	74,41 ± 0,47	91,70 ± 0,66	93,10 ± 0,54	70,45 ± 0,53	81,14 ± 0,37	00:00:00.004 ± 00:00:00.000	00:00:00.018 ± 00:00:00.000
Spam Assassin	FD	128	26	92,87 ± 1,38	85,97 ± 1,29	83,29 ± 1,90	94,01 ± 1,50	88,81 ± 0,53	00:00:00.007 ± 00:00:00.000	00:00:00.030 ± 00:00:00.000
Spam Assassin	FD	256	53	84,69 ± 0,93	93,29 ± 0,60	93,44 ± 0,56	84,38 ± 1,20	88,72 ± 0,68	00:00:00.013 ± 00:00:00.000	00:00:00.061 ± 00:00:00.002
Spam Assassin	FD	512	67	92,87 ± 0,72	87,37 ± 0,76	85,27 ± 1,00	93,94 ± 0,73	89,74 ± 0,37	00:00:00.017 ± 00:00:00.000	00:00:00.078 ± 00:00:00.002
Spam Assassin	FD	1024	58	66,43 ± 0,66	89,76 ± 0,67	92,99 ± 0,61	56,60 ± 1,41	73,28 ± 0,62	00:00:00.014 ± 00:00:00.000	00:00:00.068 ± 00:00:00.003
Spam Assassin	MI	8	3	60,22 ± 0,71	91,94 ± 0,53	95,99 ± 0,34	41,88 ± 1,26	65,40 ± 0,83	00:00:00.001 ± 00:00:00.000	00:00:00.004 ± 00:00:00.000
Spam Assassin	MI	16	6	72,16 ± 0,85	91,25 ± 0,43	92,98 ± 0,44	67,08 ± 1,40	79,19 ± 0,62	00:00:00.002 ± 00:00:00.000	00:00:00.009 ± 00:00:00.000
Spam Assassin	MI	32	18	73,40 ± 0,49	93,92 ± 0,64	95,22 ± 0,52	68,14 ± 0,72	80,86 ± 0,38	00:00:00.005 ± 00:00:00.000	00:00:00.023 ± 00:00:00.000
Spam Assassin	MI	64	33	75,88 ± 0,71	93,70 ± 0,50	94,75 ± 0,42	72,18 ± 1,10	82,85 ± 0,52	00:00:00.009 ± 00:00:00.000	00:00:00.042 ± 00:00:00.002
Spam Assassin	MI	128	58	83,95 ± 0,78	95,53 ± 0,30	95,79 ± 0,30	83,09 ± 0,81	89,16 ± 0,46	00:00:00.015 ± 00:00:00.000	00:00:00.071 ± 00:00:00.000
Spam Assassin	MI	256	85	83,36 ± 0,60	93,87 ± 0,30	94,16 ± 0,27	82,65 ± 0,76	88,16 ± 0,43	00:00:00.021 ± 00:00:00.000	00:00:00.105 ± 00:00:00.001
Spam Assassin	MI	512	96	84,10 ± 0,67	94,47 ± 0,33	94,71 ± 0,30	83,47 ± 0,78	88,85 ± 0,37	00:00:00.023 ± 00:00:00.000	00:00:00.118 ± 00:00:00.002
Spam Assassin	MI	1024	84	93,31 ± 0,66	87,03 ± 0,27	84,75 ± 0,38	94,40 ± 0,54	89,73 ± 0,28	00:00:00.021 ± 00:00:00.000	00:00:00.103 ± 00:00:00.000
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	46,53 ± 0,01	00:00:00.000 ± 00:00:00.000	00:00:00.394 ± 00:00:00.007
TREC	CHI2	16	11	61,42 ± 0,00	90,69 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	47,31 ± 0,01	00:00:00.001 ± 00:00:00.000	00:00:01.081 ± 00:00:00.015
TREC	CHI2	32	10	61,39 ± 0,00	90,23 ± 0,73	99,94 ± 0,00	0,89 ± 0,02	47,23 ± 0,01	00:00:00.001 ± 00:00:00.000	00:00:00.962 ± 00:00:00.013
TREC	CHI2	64	12	64,75 ± 7,64	88,11 ± 12,35	90,01 ± 22,49	10,90 ± 22,38	44,79 ± 5,62	00:00:00.001 ± 00:00:00.000	00:00:01.128 ± 00:00:00.024
TREC	CHI2	128	45	61,64 ± 0,01	86,11 ± 0,35	99,84 ± 0,00	1,57 ± 0,04	47,91 ± 0,03	00:00:00.009 ± 00:00:00.000	00:00:04.265 ± 00:00:00.018
TREC	CHI2	256	68	61,40 ± 0,01	91,06 ± 0,36	99,88 ± 0,01	1,95 ± 0,05	47,85 ± 0,03	00:00:00.027 ± 00:00:00.000	00:00:06.377 ± 00:00:00.021
TREC	CHI2	512	133	61,13 ± 0,02	92,49 ± 0,26	99,82 ± 0,01	3,44 ± 0,06	48,39 ± 0,06	00:00:00.168 ± 00:00:00.000	00:00:12.137 ± 00:00:00.010
TREC	CHI2	1024	47	77,23 ± 0,68	40,07 ± 0,04	2,68 ± 0,06	98,80 ± 0,04	25,76 ± 0,08	00:00:00.216 ± 00:00:00.004	00:00:03.269 ± 00:00:00.010
TREC	FD	8	6	79,70 ± 0,15	58,63 ± 0,10	33,43 ± 0,10	91,72 ± 0,09	59,49 ± 0,09	00:00:00.102 ± 00:00:00.001	00:00:00.235 ± 00:00:00.005
TREC	FD	16	7	78,40 ± 0,15	59,59 ± 0,06	36,81 ± 0,11	90,18 ± 0,10	61,11 ± 0,06	00:00:00.110 ± 00:00:00.000	00:00:00.269 ± 00:00:00.006
TREC	FD	32	14	78,75 ± 0,20	60,00 ± 0,08	37,93 ± 0,09	90,10 ± 0,11	61,79 ± 0,07	00:00:00.209 ± 00:00:00.000	00:00:00.493 ± 00:00:00.006
TREC	FD	64	7	57,36 ± 0,13	92,09 ± 0,13	97,34 ± 0,08	29,99 ± 0,40	58,49 ± 0,26	00:00:00.105 ± 00:00:00.004	00:00:00.277 ± 00:00:00.003
TREC	FD	128	22	61,09 ± 0,26	92,55 ± 0,18	96,62 ± 0,09	40,53 ± 0,64	65,45 ± 0,41	00:00:00.298 ± 00:00:00.001	00:00:00.762 ± 00:00:00.009
TREC	FD	256	24	65,83 ± 0,59	91,24 ± 0,42	94,75 ± 0,42	52,57 ± 1,51	72,08 ± 0,70	00:00:00.312 ± 00:00:00.000	00:00:00.813 ± 00:00:00.008
TREC	FD	512	51	64,19 ± 0,22	94,30 ± 0,23	96,99 ± 0,13	47,88 ± 0,42	70,25 ± 0,28	00:00:00.642 ± 00:00:00.005	00:00:01.701 ± 00:00:00.011
TREC	FD	1024	49	78,73 ± 0,50	65,33 ± 0,77	52,16 ± 2,15	86,44 ± 0,94	68,67 ± 0,77	00:00:00.610 ± 00:00:00.002	00:00:01.625 ± 00:00:00.009
TREC	MI	8	5	82,84 ± 0,15	59,65 ± 0,07	40,03 ± 0,18	91,45 ± 0,10	62,95 ± 0,09	00:00:00.078 ± 00:00:00.002	00:00:00.219 ± 00:00:00.004
TREC	MI	16	6	83,95 ± 0,07	61,50 ± 0,09	42,12 ± 0,20	91,99 ± 0,06	64,93 ± 0,11	00:00:00.087 ± 00:00:00.001	00:00:00.245 ± 00:00:00.006
TREC	MI	32	9	88,70 ± 0,08	62,64 ± 0,12	41,40 ± 0,26	94,90 ± 0,07	66,12 ± 0,15	00:00:00.123 ± 00:00:00.000	00:00:00.374 ± 00:00:00.009
TREC	MI	64	16	62,47 ± 0,37	95,25 ± 0,21	97,78 ± 0,15	43,09 ± 1,01	67,64 ± 0,58	00:00:00.212 ± 00:00:00.003	00:00:00.590 ± 00:00:00.006
TREC	MI	128	26	68,61 ± 0,91	93,17 ± 0,57	95,62 ± 0,55	57,55 ± 1,96	75,42 ± 0,89	00:00:00.338 ± 00:00:00.001	00:00:00.937 ± 00:00:00.009
TREC	MI	256	37	81,74 ± 5,28	70,02 ± 9,94	55,08 ± 18,71	84,89 ± 10,39	68,14 ± 4,81	00:00:00.465 ± 00:00:00.009	00:00:01.368 ± 00:00:00.015
TREC	MI	512	66	84,84 ± 0,79	62,76 ± 0,27	43,38 ± 0,97	92,46 ± 0,61	66,21 ± 0,36	00:00:00.805 ± 00:00:00.002	00:00:02.311 ± 00:00:00.012
TREC	MI	1024	204	85,11 ± 1,15	67,39 ± 0,77	54,23 ± 2,05	90,81 ± 1,19	71,87 ± 0,60	00:00:02.471 ± 00:00:00.011	00:00:07.260 ± 00:00:00.030
Unifei 2017	CHI2	8	8	40,96 ± 0,00	93,63 ± 0,45	99,99 ± 0,00	0,09 ± 0,00	23,89 ± 0,00	00:00:00.001 ± 00:00:00.001	00:00:02.776 ± 00:00:00.039
Unifei 2017	CHI2	16	10	40,96 ± 0,00	89,32 ± 1,48	99,98 ± 0,00	0,12 ± 0,01	23,93 ± 0,01	00:00:00.003 ± 00:00:00.000	00:00:04.403 ± 00:00:00.033
Unifei 2017	CHI2	32	20	40,95 ± 0,00	87,60 ± 1,00	99,97 ± 0,00	0,12 ± 0,01	23,92 ± 0,01	00:00:00.006 ± 00:00:00.000	00:00:08.610 ± 00:00:00.035
Unifei 2017	CHI2	64	48	40,91 ± 0,00	86,22 ± 1,51	99,96 ± 0,01	0,15 ± 0,02	23,91 ± 0,03	00:00:00.016 ± 00:00:00.000	00:00:20.394 ± 00:00:00.048
Unifei 2017	CHI2	128	86	45,77 ± 11,13	84,75 ± 6,45	89,99 ± 22,54	10,20 ± 22,56	25,97 ± 4,61	00:00:00.033 ± 00:00:00.000	00:00:36.212 ± 00:00:00.078
Unifei 2017	CHI2	256	34	63,47 ± 21,24	64,48 ± 3,10	59,70 ± 36,72	40,46 ± 36,66	32,32 ± 7,47	00:00:00.029 ± 00:00:00.000	00:00:14.285 ± 00:00:00.046
Unifei 2017	CHI2	512	45	76,59 ± 0,71	59,23 ± 0,02	1,82 ± 0,06	99,61 ± 0,01	45,20 ± 0,06	00:00:00.227 ± 00:00:00.001	00:00:17.498 ± 00:00:00.052
Unifei 2017	CHI2	1024	17	44,94 ± 0,02	99,07 ± 0,05	99,91 ± 0,00	7,12 ± 0,02	34,31 ± 0,03	00:00:00.480 ± 00:00:00.004	00:00:05.380 ± 00:00:00.028
Unifei 2017	FD	8	5	52,75 ± 0,03	71,20 ± 0,12	93,21 ± 0,03	16,74 ± 0,04	47,22 ± 0,04	00:00:00.418 ± 00:00:00.002	00:00:00.828 ± 00:00:00.021
Unifei 2017	FD	16	3	56,08 ± 0,06	80,77 ± 0,10	93,58 ± 0,03	26,90 ± 0,17	55,23 ± 0,12	00:00:00.266 ± 00:00:00.014	00:00:00.542 ± 00:00:00.019
Unifei 2017	FD	32	6	57,07 ± 0,04	79,97 ± 0,10	92,28 ± 0,04	30,74 ± 0,09	57,45 ± 0,07	00:00:00.467 ± 00:00:00.002	00:00:00.949 ± 00:00:00.023
Unifei 2017	FD	64	22	58,20 ± 0,16	70,89 ± 0,17	83,42 ± 0,17	40,25 ± 0,46	59,94 ± 0,24	00:00:01.463 ± 00:00:00.011	00:00:02.949 ± 00:00:00.019
Unifei 2017	FD	128	65	65,15 ± 0,12	81,54 ± 0,10	87,95 ± 0,08	53,09 ± 0,20	69,58 ± 0,13	00:00:04.054 ± 00:00:00.037	00:00:08.694 ± 00:00:00.028
Unifei 2017	FD	256	51	65,73 ± 0,10	79,16 ± 0,13	85,29 ± 0,14	55,67 ± 0,20	69,80 ± 0,08	00:00:03.114 ± 00:00:00.024	00:00:06.792 ± 00:00:00.023
Unifei 2017	FD	512	101	63,64 ± 0,16	83,58 ± 0,22	90,45 ± 0,14	48,47 ± 0,35	68,03 ± 0,21	00:00:06.090 ± 00:00:00.072	00:00:13.277 ± 00:00:00.058
Unifei 2017	FD	1024	105	63,53 ± 0,15	81,87 ± 0,18	89,12 ± 0,12	48,99 ± 0,27	67,73 ± 0,18	00:00:06.200 ± 00:00:00.051	00:00:13.762 ± 00:00:00.027
Unifei 2017	MI	8	4	59,90 ± 0,05	65,66 ± 0,04	68,80 ± 0,04	56,44 ± 0,06	62,32 ± 0,03	00:00:00.266 ± 00:00:00.021	00:00:00.777 ± 00:00:00.027
Unifei 2017	MI	16	7	55,21 ± 0,13	75,78 ± 0,17	89,94 ± 0,10	30,13 ± 0,37	55,49 ± 0,24	00:00:00.441 ± 00:00:00.004	00:00:01.118 ± 00:00:00.030
Unifei 2017	MI	32	5	54,42 ± 0,04	77,44 ± 0,08	92,00 ± 0,04	26,27 ± 0,05	53,49 ± 0,04	00:00:00.334 ± 00:00:00.004	00:00:00.844 ± 00:00:00.025
Unifei 2017	MI	64	9	57,53 ± 0,05	89,68 ± 0,14	96,70 ± 0,05	28,68 ± 0,08	57,80 ± 0,07	00:00:00.633 ± 00:00:00.004	00:00:01.495 ± 00:00:00.026
Unifei 2017	MI	128	18	61,51 ± 0,12	83,41 ± 0,18	91,47 ± 0,19	42,83 ± 0,42	65,07 ± 0,18	00:00:01.093 ± 00:00:00.009	00:00:02.873 ± 00:00:00.031
Unifei 2017	MI	256	31	68,54 ± 0,21	79,72 ± 0,25	84,33 ± 0,36	61,40 ± 0,54	72,49 ± 0,14	00:00:01.876 ± 00:00:00.015	00:00:04.436 ± 00:00:00.024
Unifei 2017	MI	512	125	72,05 ± 0,35	80,48 ± 0,29	83,52 ± 0,45	67,69 ± 0,70	75,44 ± 0,17	00:00:07.338 ± 00:00:00.064	00:00:16.914 ± 00:00:00.034
Unifei 2017	MI	1024	208	67,47 ± 0,17	85,68 ± 0,21	90,53 ± 0,16	56,49 ± 0,38	72,69 ± 0,20	00:00:12.250 ± 00:00:00.124	00:00:29.795 ± 00:00:00.074
Unifei 2018	CHI2	8	4	99,84 ± 0,03	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	44,60 ± 0,00	00:00:00.004 ± 00:00:00.002	00:00:01.594 ± 00:00:00.053
Unifei 2018	CHI2	16	16	41,64 ± 0,01	99,01 ± 0,15	99,98 ± 0,00	1,31 ± 0,02	25,81 ± 0,03	00:00:00.013 ± 00:00:00.000	00:00:05.633 ± 00:00:00.084
Unifei 2018	CHI2	32	28	41,62 ± 0,01	98,80 ± 0,11	99,97 ± 0,00	1,74 ± 0,03	26,23 ± 0,04	00:00:00.037 ± 00:00:00.000	00:00:10.200 ± 00:00:00.054
Unifei 2018	CHI2	64	42	41,59 ± 0,00	97,81 ± 0,24	99,92 ± 0,01	2,34 ± 0,03	26,80 ± 0,03	00:00:00.077 ± 00:00:00.003	00:00:15.469 ± 00:00:00.093
Unifei 2018	CHI2	128	69	97,23 ± 0,06	59,60 ± 0,01	7,41 ± 0,02	99,85 ± 0,00	48,94 ± 0,02	00:00:00.395 ± 00:00:00.001	00:00:26.892 ± 00:00:00.100
Unifei 2018	CHI2	256	116	94,68 ± 0,10	60,07 ± 0,01	7,17 ± 0,06	99,71 ± 0,00	49,30 ± 0,05	00:00:00.740 ± 00:00:00.002	00:00:44.064 ± 00:00:00.063
Unifei 2018	CHI2	512	121	95,60 ± 0,11	57,06 ± 0,02	14,65 ± 0,03	99,41 ± 0,01	50,50 ± 0,03	00:00:02.913 ± 00:00:00.101	00:00:42.871 ± 00:00:00.063
Unifei 2018	CHI2	1024	152	93,55 ± 0,07	54,60 ± 0,03	9,73 ± 0,02	99,39 ± 0,01	45,22 ± 0,04	00:00:05.652 ± 00:00:00.008	00:00:46.847 ± 00:00:00.168
Unifei 2018	FD	8	5	52,30 ± 0,04	70,83 ± 0,18	93,08 ± 0,04	16,53 ± 0,05	46,72 ± 0,05	00:00:00.443 ± 00:00:00.002	00:00:00.861 ± 00:00:00.026
Unifei 2018	FD	16	3	56,77 ± 0,05	79,66 ± 0,14	91,82 ± 0,08	31,43 ± 0,06	57,49 ± 0,04	00:00:00.300 ± 00:00:00.020	00:00:00.551 ± 00:00:00.023
Unifei 2018	FD	32	8	55,68 ± 0,04	80,04 ± 0,10	92,84 ± 0,05	27,99 ± 0,13	55,36 ± 0,08	00:00:00.626 ± 00:00:00.006	00:00:01.203 ± 00:00:00.036
Unifei 2018	FD	64	11	54,54 ± 0,03	76,38 ± 0,09	91,94 ± 0,05	25,38 ± 0,04	53,08 ± 0,04	00:00:00.761 ± 00:00:00.032	00:00:01.590 ± 00:00:00.029
Unifei 2018	FD	128	24	59,94 ± 0,16	78,80 ± 0,30	88,25 ± 0,14	42,54 ± 0,36	63,22 ± 0,25	00:00:01.582 ± 00:00:00.071	00:00:03.276 ± 00:00:00.043
Unifei 2018	FD	256	60	63,01 ± 0,07	81,83 ± 0,14	88,77 ± 0,12	49,24 ± 0,16	67,51 ± 0,07	00:00:03.813 ± 00:00:00.058	00:00:08.243 ± 00:00:00.140
Unifei 2018	FD	512	79	66,85 ± 1,10	83,62 ± 0,26	88,50 ± 0,39	57,19 ± 2,37	71,96 ± 1,17	00:00:04.986 ± 00:00:00.061	00:00:10.410 ± 00:00:00.082
Unifei 2018	FD	1024	77	76,96 ± 0,33	66,67 ± 0,21	57,32 ± 0,63	83,26 ± 0,47	69,92 ± 0,17	00:00:04.839 ± 00:00:00.074	00:00:10.189 ± 00:00:00.168
Unifei 2018	MI	8	3	70,62 ± 3,43	50,92 ± 0,46	9,95 ± 1,83	95,92 ± 0,19	41,62 ± 1,67	00:00:00.330 ± 00:00:00.004	00:00:00.964 ± 00:00:00.041
Unifei 2018	MI	16	3	64,24 ± 0,03	80,42 ± 0,09	86,15 ± 0,08	54,25 ± 0,06	69,09 ± 0,03	00:00:00.291 ± 00:00:00.002	00:00:00.657 ± 00:00:00.032
Unifei 2018	MI	32	7	55,93 ± 0,06	77,41 ± 0,09	90,36 ± 0,05	31,69 ± 0,12	56,78 ± 0,08	00:00:00.530 ± 00:00:00.005	00:00:01.230 ± 00:00:00.032
Unifei 2018	MI	64	10	57,68 ± 0,04	86,81 ± 0,09	95,15 ± 0,04	31,40 ± 0,07	58,86 ± 0,04	00:00:00.687 ± 00:00:00.019	00:00:01.533 ± 00:00:00.035
Unifei 2018	MI	128	30	60,27 ± 0,09	87,89 ± 0,06	94,41 ± 0,04	39,45 ± 0,26	63,88 ± 0,16	00:00:01.946 ± 00:00:00.002	00:00:04.414 ± 00:00:00.034
Unifei 2018	MI	256	49	63,36 ± 0,06	81,26 ± 0,09	88,10 ± 0,09	50,32 ± 0,13	67,86 ± 0,04	00:00:03.080 ± 00:00:00.053	00:00:06.769 ± 00:00:00.128
Unifei 2018	MI	512	118	62,86 ± 0,15	83,76 ± 0,16	90,47 ± 0,16	47,89 ± 0,38	67,47 ± 0,18	00:00:07.288 ± 00:00:00.050	00:00:15.927 ± 00:00:00.160
Unifei 2018	MI	1024	155	90,78 ± 0,14	66,02 ± 0,22	49,82 ± 0,49	95,07 ± 0,08	71,21 ± 0,28	00:00:09.443 ± 00:00:00.064	00:00:22.212 ± 00:00:00.127

Tabela 19: Resultados do classificador AODE.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.006 ± 00:00:00.011	00:00:00.012 ± 00:00:00.009
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	77,42 ± 0,04	00:00:00.000 ± 00:00:00.000	00:00:00.006 ± 00:00:00.000
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	78,40 ± 0,03	00:00:00.001 ± 00:00:00.000	00:00:00.005 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,30 ± 0,14	99,94 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.003 ± 00:00:00.001	00:00:00.011 ± 00:00:00.005
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,47 ± 0,09	99,92 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.002 ± 00:00:00.000	00:00:00.010 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,75 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:00.005 ± 00:00:00.000	00:00:00.022 ± 00:00:00.000
Ling Spam	CHI2	512	14	68,47 ± 0,14	90,09 ± 0,48	97,11 ± 0,16	37,01 ± 0,47	68,75 ± 0,21	00:00:00.008 ± 00:00:00.000	00:00:00.021 ± 00:00:00.000
Ling Spam	CHI2	1024	242	72,69 ± 0,26	88,31 ± 0,31	95,88 ± 0,16	46,32 ± 0,51	73,88 ± 0,27	00:00:00.426 ± 00:00:00.006	00:00:02.145 ± 00:00:00.001
Ling Spam	FD	8	4	96,99 ± 0,25	84,74 ± 0,18	80,80 ± 0,27	97,70 ± 0,21	89,51 ± 0,10	00:00:00.006 ± 00:00:00.000	00:00:00.007 ± 00:00:00.000
Ling Spam	FD	16	4	97,08 ± 0,25	84,74 ± 0,18	80,79 ± 0,28	97,77 ± 0,20	89,55 ± 0,07	00:00:00.006 ± 00:00:00.000	00:00:00.008 ± 00:00:00.000
Ling Spam	FD	32	10	98,24 ± 0,17	91,50 ± 0,16	90,01 ± 0,17	98,53 ± 0,14	94,44 ± 0,10	00:00:00.015 ± 00:00:00.000	00:00:00.014 ± 00:00:00.000
Ling Spam	FD	64	19	98,81 ± 0,19	93,56 ± 0,14	92,57 ± 0,18	98,98 ± 0,16	95,90 ± 0,14	00:00:00.026 ± 00:00:00.000	00:00:00.029 ± 00:00:00.001
Ling Spam	FD	128	38	99,08 ± 0,12	94,96 ± 0,21	94,26 ± 0,26	99,19 ± 0,11	96,83 ± 0,10	00:00:00.058 ± 00:00:00.003	00:00:00.075 ± 00:00:00.000
Ling Spam	FD	256	58	99,51 ± 0,12	96,27 ± 0,16	95,78 ± 0,19	99,56 ± 0,11	97,75 ± 0,08	00:00:00.097 ± 00:00:00.000	00:00:00.157 ± 00:00:00.003
Ling Spam	FD	512	103	99,80 ± 0,12	97,47 ± 0,10	97,17 ± 0,11	99,82 ± 0,11	98,55 ± 0,05	00:00:00.236 ± 00:00:00.000	00:00:00.418 ± 00:00:00.000
Ling Spam	FD	1024	121	99,72 ± 0,16	96,92 ± 0,12	96,54 ± 0,14	99,76 ± 0,15	98,21 ± 0,11	00:00:00.293 ± 00:00:00.000	00:00:00.560 ± 00:00:00.000
Ling Spam	MI	8	6	98,22 ± 0,23	85,00 ± 0,14	81,00 ± 0,19	98,65 ± 0,18	90,11 ± 0,08	00:00:00.008 ± 00:00:00.000	00:00:00.010 ± 00:00:00.001
Ling Spam	MI	16	12	98,80 ± 0,23	91,19 ± 0,18	89,56 ± 0,23	99,00 ± 0,19	94,47 ± 0,10	00:00:00.017 ± 00:00:00.000	00:00:00.016 ± 00:00:00.000
Ling Spam	MI	32	19	91,43 ± 0,28	97,85 ± 0,05	97,81 ± 0,06	91,61 ± 0,25	94,57 ± 0,15	00:00:00.025 ± 00:00:00.000	00:00:00.028 ± 00:00:00.000
Ling Spam	MI	64	34	95,87 ± 0,14	98,20 ± 0,13	98,08 ± 0,14	96,13 ± 0,12	97,06 ± 0,09	00:00:00.047 ± 00:00:00.000	00:00:00.063 ± 00:00:00.000
Ling Spam	MI	128	52	95,93 ± 0,26	98,16 ± 0,10	98,03 ± 0,10	96,19 ± 0,23	97,07 ± 0,13	00:00:00.079 ± 00:00:00.000	00:00:00.127 ± 00:00:00.001
Ling Spam	MI	256	68	99,42 ± 0,17	96,50 ± 0,17	96,06 ± 0,19	99,49 ± 0,15	97,85 ± 0,10	00:00:00.114 ± 00:00:00.000	00:00:00.200 ± 00:00:00.000
Ling Spam	MI	512	139	97,34 ± 0,21	99,17 ± 0,10	99,11 ± 0,11	97,52 ± 0,19	98,28 ± 0,12	00:00:00.361 ± 00:00:00.000	00:00:00.720 ± 00:00:00.002
Ling Spam	MI	1024	157	97,39 ± 0,15	99,20 ± 0,06	99,14 ± 0,06	97,57 ± 0,13	98,32 ± 0,09	00:00:00.434 ± 00:00:00.001	00:00:00.907 ± 00:00:00.001
Spam Assassin	CHI2	8	4	65,26 ± 0,53	79,23 ± 0,77	82,63 ± 0,67	60,10 ± 0,61	70,52 ± 0,29	00:00:00.002 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Spam Assassin	CHI2	16	7	67,86 ± 0,46	80,15 ± 0,60	82,28 ± 0,53	64,75 ± 0,72	72,94 ± 0,46	00:00:00.003 ± 00:00:00.000	00:00:00.004 ± 00:00:00.000
Spam Assassin	CHI2	32	14	69,19 ± 0,51	81,34 ± 0,41	83,14 ± 0,39	66,48 ± 0,78	74,28 ± 0,37	00:00:00.005 ± 00:00:00.000	00:00:00.008 ± 00:00:00.000
Spam Assassin	CHI2	64	20	69,12 ± 0,63	81,30 ± 0,38	82,81 ± 0,43	66,88 ± 0,79	74,31 ± 0,43	00:00:00.008 ± 00:00:00.000	00:00:00.013 ± 00:00:00.000
Spam Assassin	CHI2	128	25	73,21 ± 0,47	82,07 ± 0,88	82,52 ± 0,79	72,61 ± 0,68	77,30 ± 0,49	00:00:00.010 ± 00:00:00.000	00:00:00.017 ± 00:00:00.000
Spam Assassin	CHI2	256	29	77,84 ± 0,90	86,11 ± 0,41	86,26 ± 0,60	77,59 ± 1,12	81,72 ± 0,51	00:00:00.013 ± 00:00:00.000	00:00:00.021 ± 00:00:00.000
Spam Assassin	CHI2	512	7	62,59 ± 0,35	92,90 ± 0,43	96,00 ± 0,25	47,70 ± 0,33	69,11 ± 0,25	00:00:00.003 ± 00:00:00.000	00:00:00.004 ± 00:00:00.000
Spam Assassin	CHI2	1024	20	66,93 ± 1,70	81,07 ± 1,61	84,38 ± 2,37	61,19 ± 3,67	71,96 ± 1,06	00:00:00.007 ± 00:00:00.000	00:00:00.012 ± 00:00:00.000
Spam Assassin	FD	8	5	67,87 ± 0,46	81,96 ± 1,10	84,82 ± 1,35	63,11 ± 1,17	73,25 ± 0,28	00:00:00.003 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Spam Assassin	FD	16	6	80,77 ± 0,53	88,69 ± 0,56	88,87 ± 0,64	80,46 ± 0,76	84,49 ± 0,40	00:00:00.003 ± 00:00:00.000	00:00:00.004 ± 00:00:00.000
Spam Assassin	FD	32	9	82,93 ± 0,90	93,68 ± 0,26	94,00 ± 0,31	82,11 ± 1,23	87,80 ± 0,61	00:00:00.005 ± 00:00:00.000	00:00:00.005 ± 00:00:00.000
Spam Assassin	FD	64	16	87,64 ± 0,60	94,42 ± 0,39	94,40 ± 0,32	87,70 ± 0,75	90,92 ± 0,47	00:00:00.009 ± 00:00:00.000	00:00:00.010 ± 00:00:00.000
Spam Assassin	FD	128	26	91,37 ± 0,75	94,45 ± 0,44	94,16 ± 0,43	91,78 ± 0,82	92,92 ± 0,42	00:00:00.014 ± 00:00:00.000	00:00:00.018 ± 00:00:00.000
Spam Assassin	FD	256	53	91,90 ± 0,41	97,33 ± 0,36	97,27 ± 0,35	92,08 ± 0,42	94,57 ± 0,24	00:00:00.035 ± 00:00:00.006	00:00:00.056 ± 00:00:00.000
Spam Assassin	FD	512	67	90,79 ± 0,31	96,78 ± 0,39	96,73 ± 0,37	90,94 ± 0,31	93,72 ± 0,21	00:00:00.045 ± 00:00:00.000	00:00:00.084 ± 00:00:00.000
Spam Assassin	FD	1024	58	80,90 ± 0,48	95,93 ± 0,32	96,37 ± 0,25	78,98 ± 0,81	87,27 ± 0,45	00:00:00.034 ± 00:00:00.000	00:00:00.065 ± 00:00:00.000
Spam Assassin	MI	8	3	80,54 ± 0,33	85,98 ± 0,43	85,56 ± 0,59	81,05 ± 0,56	83,22 ± 0,18	00:00:00.002 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Spam Assassin	MI	16	6	87,28 ± 0,79	89,45 ± 0,58	88,65 ± 0,67	88,15 ± 0,79	88,39 ± 0,27	00:00:00.003 ± 00:00:00.000	00:00:00.004 ± 00:00:00.000
Spam Assassin	MI	32	18	87,76 ± 0,51	95,35 ± 0,41	95,37 ± 0,39	87,71 ± 0,64	91,38 ± 0,37	00:00:00.010 ± 00:00:00.000	00:00:00.011 ± 00:00:00.000
Spam Assassin	MI	64	33	89,11 ± 0,52	97,14 ± 0,28	97,17 ± 0,25	89,03 ± 0,63	92,94 ± 0,35	00:00:00.019 ± 00:00:00.000	00:00:00.026 ± 00:00:00.000
Spam Assassin	MI	128	58	91,31 ± 0,42	98,31 ± 0,18	98,30 ± 0,17	91,37 ± 0,44	94,70 ± 0,22	00:00:00.037 ± 00:00:00.000	00:00:00.066 ± 00:00:00.000
Spam Assassin	MI	256	85	91,15 ± 0,42	98,17 ± 0,19	98,16 ± 0,18	91,20 ± 0,46	94,54 ± 0,24	00:00:00.066 ± 00:00:00.000	00:00:00.127 ± 00:00:00.001
Spam Assassin	MI	512	96	91,97 ± 0,52	97,32 ± 0,22	97,26 ± 0,21	92,16 ± 0,54	94,61 ± 0,27	00:00:00.079 ± 00:00:00.000	00:00:00.156 ± 00:00:00.000
Spam Assassin	MI	1024	84	92,25 ± 0,57	96,31 ± 0,35	96,17 ± 0,33	92,54 ± 0,58	94,28 ± 0,36	00:00:00.062 ± 00:00:00.001	00:00:00.123 ± 00:00:00.000
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	46,53 ± 0,01	00:00:00.000 ± 00:00:00.000	00:00:00.161 ± 00:00:00.002
TREC	CHI2	16	11	61,42 ± 0,00	90,69 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	47,31 ± 0,01	00:00:00.002 ± 00:00:00.000	00:00:00.312 ± 00:00:00.005
TREC	CHI2	32	10	61,39 ± 0,00	90,24 ± 0,73	99,94 ± 0,00	0,89 ± 0,02	47,23 ± 0,01	00:00:00.002 ± 00:00:00.000	00:00:00.286 ± 00:00:00.003
TREC	CHI2	64	12	61,37 ± 0,00	93,62 ± 0,42	99,96 ± 0,00	1,01 ± 0,01	47,28 ± 0,01	00:00:00.003 ± 00:00:00.000	00:00:00.337 ± 00:00:00.000
TREC	CHI2	128	45	61,70 ± 0,01	87,84 ± 0,20	99,84 ± 0,00	1,82 ± 0,02	48,13 ± 0,02	00:00:00.020 ± 00:00:00.000	00:00:01.956 ± 00:00:00.018
TREC	CHI2	256	68	91,06 ± 0,23	39,38 ± 0,01	1,68 ± 0,02	99,74 ± 0,01	24,06 ± 0,02	00:00:00.072 ± 00:00:00.003	00:00:03.919 ± 00:00:00.016
TREC	CHI2	512	133	70,77 ± 10,93	77,36 ± 18,29	70,81 ± 33,40	32,74 ± 33,02	42,17 ± 7,77	00:00:00.696 ± 00:00:00.056	00:00:12.317 ± 00:00:00.007
TREC	CHI2	1024	47	61,72 ± 0,04	76,57 ± 0,30	98,55 ± 0,02	7,18 ± 0,07	50,98 ± 0,06	00:00:00.531 ± 00:00:00.012	00:00:01.681 ± 00:00:00.003
TREC	FD	8	6	74,81 ± 0,15	78,83 ± 0,36	79,57 ± 0,50	73,95 ± 0,28	76,71 ± 0,16	00:00:00.353 ± 00:00:00.004	00:00:00.147 ± 00:00:00.010
TREC	FD	16	7	78,27 ± 0,23	77,96 ± 0,36	76,82 ± 0,47	79,36 ± 0,26	78,10 ± 0,27	00:00:00.393 ± 00:00:00.010	00:00:00.157 ± 00:00:00.000
TREC	FD	32	14	81,70 ± 0,23	85,09 ± 0,16	85,23 ± 0,20	81,52 ± 0,28	83,34 ± 0,10	00:00:00.772 ± 00:00:00.013	00:00:00.295 ± 00:00:00.010
TREC	FD	64	7	68,64 ± 0,09	87,91 ± 0,17	91,54 ± 0,17	59,54 ± 0,17	74,66 ± 0,07	00:00:00.289 ± 00:00:00.001	00:00:00.146 ± 00:00:00.000
TREC	FD	128	22	77,34 ± 0,20	90,16 ± 0,14	91,64 ± 0,12	74,06 ± 0,24	82,58 ± 0,13	00:00:01.014 ± 00:00:00.008	00:00:00.488 ± 00:00:00.013
TREC	FD	256	24	75,35 ± 0,13	89,67 ± 0,17	91,49 ± 0,16	71,16 ± 0,18	80,97 ± 0,10	00:00:00.998 ± 00:00:00.008	00:00:00.517 ± 00:00:00.001
TREC	FD	512	51	78,72 ± 0,11	92,84 ± 0,16	93,95 ± 0,15	75,54 ± 0,19	84,46 ± 0,10	00:00:02.208 ± 00:00:00.018	00:00:01.900 ± 00:00:00.200
TREC	FD	1024	49	75,64 ± 0,10	90,34 ± 0,09	92,03 ± 0,08	71,55 ± 0,11	81,41 ± 0,07	00:00:02.152 ± 00:00:00.034	00:00:01.641 ± 00:00:00.020
TREC	MI	8	5	80,53 ± 0,84	65,41 ± 0,34	55,88 ± 1,07	86,03 ± 1,01	70,07 ± 0,19	00:00:00.201 ± 00:00:00.006	00:00:00.123 ± 00:00:00.008
TREC	MI	16	6	81,56 ± 0,26	67,16 ± 0,12	57,17 ± 0,24	87,14 ± 0,25	71,55 ± 0,11	00:00:00.254 ± 00:00:00.007	00:00:00.132 ± 00:00:00.002
TREC	MI	32	9	65,55 ± 0,09	90,80 ± 0,15	94,54 ± 0,12	52,00 ± 0,17	71,68 ± 0,08	00:00:00.353 ± 00:00:00.004	00:00:00.176 ± 00:00:00.008
TREC	MI	64	16	73,60 ± 0,09	89,67 ± 0,10	91,91 ± 0,09	68,06 ± 0,11	79,53 ± 0,05	00:00:00.597 ± 00:00:00.003	00:00:00.304 ± 00:00:00.007
TREC	MI	128	26	77,84 ± 0,08	90,50 ± 0,15	91,91 ± 0,14	74,65 ± 0,18	83,03 ± 0,09	00:00:01.091 ± 00:00:00.009	00:00:00.592 ± 00:00:00.013
TREC	MI	256	37	75,20 ± 0,10	90,41 ± 0,16	92,27 ± 0,14	70,56 ± 0,16	81,03 ± 0,05	00:00:01.410 ± 00:00:00.025	00:00:00.971 ± 00:00:00.014
TREC	MI	512	66	77,33 ± 0,06	91,44 ± 0,12	92,89 ± 0,10	73,61 ± 0,10	82,96 ± 0,06	00:00:02.778 ± 00:00:00.030	00:00:02.647 ± 00:00:00.020
TREC	MI	1024	204	82,83 ± 0,10	93,91 ± 0,09	94,53 ± 0,09	81,16 ± 0,11	87,67 ± 0,07	00:00:13.285 ± 00:00:00.050	00:00:30.387 ± 00:00:00.568
Unifei 2017	CHI2	8	8	40,96 ± 0,00	95,70 ± 0,36	99,99 ± 0,00	0,09 ± 0,00	23,89 ± 0,00	00:00:00.011 ± 00:00:00.017	00:00:01.026 ± 00:00:00.028
Unifei 2017	CHI2	16	10	40,96 ± 0,00	90,65 ± 2,10	99,98 ± 0,01	0,12 ± 0,01	23,93 ± 0,01	00:00:00.008 ± 00:00:00.001	00:00:01.175 ± 00:00:00.008
Unifei 2017	CHI2	32	20	40,95 ± 0,00	87,22 ± 1,87	99,97 ± 0,01	0,15 ± 0,01	23,95 ± 0,01	00:00:00.011 ± 00:00:00.000	00:00:02.424 ± 00:00:00.040
Unifei 2017	CHI2	64	48	40,91 ± 0,00	86,38 ± 1,37	99,96 ± 0,01	0,17 ± 0,01	23,93 ± 0,01	00:00:00.031 ± 00:00:00.001	00:00:08.626 ± 00:00:00.036
Unifei 2017	CHI2	128	86	40,85 ± 0,00	87,04 ± 0,36	99,94 ± 0,00	0,26 ± 0,01	23,98 ± 0,01	00:00:00.080 ± 00:00:00.002	00:00:24.079 ± 00:00:00.033
Unifei 2017	CHI2	256	34	40,55 ± 0,00	68,17 ± 0,35	99,46 ± 0,01	0,79 ± 0,01	24,26 ± 0,01	00:00:00.058 ± 00:00:00.002	00:00:04.993 ± 00:00:00.053
Unifei 2017	CHI2	512	45	41,31 ± 0,01	64,25 ± 0,10	97,00 ± 0,01	3,76 ± 0,01	28,01 ± 0,01	00:00:00.494 ± 00:00:00.004	00:00:07.351 ± 00:00:00.044
Unifei 2017	CHI2	1024	17	44,98 ± 0,02	98,98 ± 0,04	99,90 ± 0,00	7,25 ± 0,02	34,44 ± 0,03	00:00:01.085 ± 00:00:00.020	00:00:01.526 ± 00:00:00.012
Unifei 2017	FD	8	5	87,91 ± 0,13	86,99 ± 0,11	86,78 ± 0,15	88,10 ± 0,16	87,44 ± 0,04	00:00:02.346 ± 00:00:00.036	00:00:00.762 ± 00:00:00.017
Unifei 2017	FD	16	3	67,75 ± 0,07	87,59 ± 0,26	92,00 ± 0,23	56,32 ± 0,23	73,29 ± 0,05	00:00:01.459 ± 00:00:00.083	00:00:00.439 ± 00:00:00.026
Unifei 2017	FD	32	6	76,68 ± 0,04	88,98 ± 0,07	91,01 ± 0,07	72,40 ± 0,07	81,53 ± 0,03	00:00:03.123 ± 00:00:00.057	00:00:00.741 ± 00:00:00.032
Unifei 2017	FD	64	22	86,92 ± 0,22	94,72 ± 0,08	95,21 ± 0,09	85,72 ± 0,27	90,43 ± 0,12	00:00:09.673 ± 00:00:00.270	00:00:02.594 ± 00:00:00.053
Unifei 2017	FD	128	65	88,69 ± 0,11	96,83 ± 0,05	97,12 ± 0,05	87,65 ± 0,15	92,37 ± 0,06	00:00:29.522 ± 00:00:00.336	00:00:17.162 ± 00:00:00.255
Unifei 2017	FD	256	51	85,70 ± 0,05	95,72 ± 0,09	96,24 ± 0,08	83,99 ± 0,06	90,06 ± 0,04	00:00:21.194 ± 00:00:00.369	00:00:09.505 ± 00:00:00.081
Unifei 2017	FD	512	101	87,69 ± 0,03	96,95 ± 0,05	97,27 ± 0,05	86,39 ± 0,04	91,80 ± 0,03	00:00:43.768 ± 00:00:00.682	00:00:32.986 ± 00:00:00.359
Unifei 2017	FD	1024	105	81,18 ± 0,07	96,78 ± 0,06	97,42 ± 0,05	77,48 ± 0,11	87,31 ± 0,05	00:00:41.948 ± 00:00:00.713	00:00:31.946 ± 00:00:00.187
Unifei 2017	MI	8	4	63,56 ± 0,05	70,42 ± 0,10	73,24 ± 0,16	60,27 ± 0,07	66,46 ± 0,05	00:00:01.257 ± 00:00:00.008	00:00:00.551 ± 00:00:00.034
Unifei 2017	MI	16	7	59,90 ± 0,03	89,55 ± 0,09	95,26 ± 0,05	38,94 ± 0,05	63,70 ± 0,03	00:00:01.804 ± 00:00:00.026	00:00:00.648 ± 00:00:00.030
Unifei 2017	MI	32	5	60,25 ± 0,03	92,59 ± 0,10	96,74 ± 0,05	38,92 ± 0,06	64,32 ± 0,04	00:00:01.087 ± 00:00:00.016	00:00:00.526 ± 00:00:00.026
Unifei 2017	MI	64	9	72,80 ± 0,08	88,50 ± 0,15	91,43 ± 0,15	65,87 ± 0,17	78,29 ± 0,04	00:00:02.419 ± 00:00:00.029	00:00:00.857 ± 00:00:00.039
Unifei 2017	MI	128	18	69,11 ± 0,04	91,58 ± 0,15	94,69 ± 0,11	57,72 ± 0,11	75,35 ± 0,03	00:00:04.182 ± 00:00:00.071	00:00:01.422 ± 00:00:00.037
Unifei 2017	MI	256	31	79,09 ± 0,10	92,44 ± 0,11	93,83 ± 0,10	75,26 ± 0,15	84,40 ± 0,06	00:00:08.596 ± 00:00:00.140	00:00:03.126 ± 00:00:00.032
Unifei 2017	MI	512	125	84,90 ± 0,04	96,40 ± 0,12	96,89 ± 0,11	82,82 ± 0,07	89,79 ± 0,05	00:00:44.757 ± 00:00:00.619	00:00:42.276 ± 00:00:00.293
Unifei 2017	MI	1024	208	86,14 ± 0,05	97,18 ± 0,03	97,55 ± 0,03	84,35 ± 0,06	90,90 ± 0,04	00:01:23.887 ± 00:00:00.491	00:02:11.424 ± 00:00:08.876
Unifei 2018	CHI2	8	4	99,78 ± 0,03	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	44,61 ± 0,00	00:00:00.022 ± 00:00:00.028	00:00:00.670 ± 00:00:00.030
Unifei 2018	CHI2	16	16	99,62 ± 0,05	59,17 ± 0,00	2,03 ± 0,01	99,99 ± 0,00	45,27 ± 0,00	00:00:00.024 ± 00:00:00.000	00:00:01.841 ± 00:00:00.053
Unifei 2018	CHI2	32	28	99,29 ± 0,06	59,46 ± 0,00	2,70 ± 0,01	99,99 ± 0,00	46,01 ± 0,00	00:00:00.095 ± 00:00:00.005	00:00:03.704 ± 00:00:00.029
Unifei 2018	CHI2	64	42	41,65 ± 0,00	98,75 ± 0,08	99,95 ± 0,00	2,56 ± 0,01	27,06 ± 0,01	00:00:00.220 ± 00:00:00.002	00:00:06.882 ± 00:00:00.041
Unifei 2018	CHI2	128	69	98,14 ± 0,22	59,89 ± 0,01	8,48 ± 0,02	99,88 ± 0,01	49,85 ± 0,01	00:00:01.202 ± 00:00:00.010	00:00:16.403 ± 00:00:00.043
Unifei 2018	CHI2	256	116	97,88 ± 0,04	60,47 ± 0,00	8,55 ± 0,02	99,87 ± 0,00	50,50 ± 0,01	00:00:02.762 ± 00:00:00.071	00:00:40.218 ± 00:00:00.085
Unifei 2018	CHI2	512	121	55,42 ± 10,30	91,08 ± 8,27	90,93 ± 18,20	24,51 ± 18,82	46,87 ± 1,69	00:00:10.915 ± 00:00:00.037	00:00:40.738 ± 00:00:00.079
Unifei 2018	CHI2	1024	152	53,88 ± 0,02	90,41 ± 0,08	97,24 ± 0,02	23,80 ± 0,05	52,81 ± 0,04	00:00:22.591 ± 00:00:00.137	00:00:53.661 ± 00:00:00.119
Unifei 2018	FD	8	5	95,97 ± 0,05	94,23 ± 0,07	94,02 ± 0,08	96,12 ± 0,05	95,07 ± 0,04	00:00:02.406 ± 00:00:00.028	00:00:00.747 ± 00:00:00.040
Unifei 2018	FD	16	3	71,46 ± 0,10	87,87 ± 0,17	90,93 ± 0,16	64,38 ± 0,19	77,14 ± 0,05	00:00:01.468 ± 00:00:00.025	00:00:00.479 ± 00:00:00.036
Unifei 2018	FD	32	8	83,93 ± 0,08	92,18 ± 0,09	92,80 ± 0,09	82,69 ± 0,11	87,65 ± 0,04	00:00:04.140 ± 00:00:00.080	00:00:00.887 ± 00:00:00.043
Unifei 2018	FD	64	11	73,26 ± 0,05	91,57 ± 0,12	93,69 ± 0,10	66,69 ± 0,10	79,67 ± 0,03	00:00:04.327 ± 00:00:00.104	00:00:00.909 ± 00:00:00.038
Unifei 2018	FD	128	24	82,73 ± 0,08	96,11 ± 0,05	96,66 ± 0,04	80,35 ± 0,11	88,33 ± 0,05	00:00:10.066 ± 00:00:00.215	00:00:02.458 ± 00:00:00.077
Unifei 2018	FD	256	60	86,68 ± 0,05	97,58 ± 0,04	97,83 ± 0,04	85,36 ± 0,05	91,48 ± 0,02	00:00:24.708 ± 00:00:00.378	00:00:12.270 ± 00:00:00.086
Unifei 2018	FD	512	79	87,38 ± 0,05	97,80 ± 0,03	98,01 ± 0,02	86,22 ± 0,05	92,01 ± 0,03	00:00:33.002 ± 00:00:00.290	00:00:20.629 ± 00:00:00.308
Unifei 2018	FD	1024	77	84,22 ± 0,05	96,52 ± 0,04	96,96 ± 0,03	82,28 ± 0,06	89,48 ± 0,03	00:00:27.740 ± 00:00:00.355	00:00:18.416 ± 00:00:00.243
Unifei 2018	MI	8	3	96,25 ± 0,02	72,05 ± 0,04	63,19 ± 0,06	97,47 ± 0,01	79,53 ± 0,03	00:00:00.964 ± 00:00:00.011	00:00:00.622 ± 00:00:00.038
Unifei 2018	MI	16	3	93,26 ± 0,05	88,01 ± 0,06	86,58 ± 0,07	94,03 ± 0,05	90,37 ± 0,03	00:00:01.471 ± 00:00:00.019	00:00:00.568 ± 00:00:00.014
Unifei 2018	MI	32	7	93,54 ± 0,11	90,16 ± 0,05	89,30 ± 0,06	94,09 ± 0,10	91,73 ± 0,03	00:00:02.683 ± 00:00:00.129	00:00:00.836 ± 00:00:00.039
Unifei 2018	MI	64	10	68,33 ± 0,04	94,65 ± 0,09	96,78 ± 0,06	55,93 ± 0,08	75,17 ± 0,04	00:00:02.850 ± 00:00:00.057	00:00:00.819 ± 00:00:00.046
Unifei 2018	MI	128	30	79,49 ± 0,09	95,33 ± 0,05	96,18 ± 0,04	75,86 ± 0,12	85,75 ± 0,06	00:00:09.724 ± 00:00:00.249	00:00:03.267 ± 00:00:00.041
Unifei 2018	MI	256	49	81,55 ± 0,07	96,57 ± 0,05	97,14 ± 0,04	78,56 ± 0,12	87,64 ± 0,06	00:00:17.126 ± 00:00:00.219	00:00:07.762 ± 00:00:00.188
Unifei 2018	MI	512	118	86,19 ± 0,04	97,79 ± 0,03	98,03 ± 0,03	84,69 ± 0,06	91,24 ± 0,03	00:00:45.311 ± 00:00:00.376	00:00:40.673 ± 00:00:00.258
Unifei 2018	MI	1024	155	86,25 ± 0,09	97,84 ± 0,02	98,09 ± 0,02	84,76 ± 0,11	91,30 ± 0,05	00:00:59.219 ± 00:00:00.393	00:01:03.789 ± 00:00:00.374

Tabela 20: Resultados do classificador SLP.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.010 ± 00:00:00.004	00:00:00.006 ± 00:00:00.004
Ling Spam	CHI2	16	1	83,00 ± 0,36	86,78 ± 21,82	99,93 ± 0,02	8,90 ± 2,25	77,00 ± 0,93	00:00:00.016 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	32	1	82,30 ± 0,85	78,69 ± 29,67	99,94 ± 0,03	14,03 ± 5,29	76,96 ± 2,19	00:00:00.031 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,30 ± 0,14	99,94 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.066 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,47 ± 0,09	99,92 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.127 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,75 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:00.322 ± 00:00:00.003	00:00:00.004 ± 00:00:00.000
Ling Spam	CHI2	512	14	73,89 ± 10,07	81,31 ± 14,88	78,54 ± 28,80	46,13 ± 21,77	58,12 ± 12,91	00:00:00.563 ± 00:00:00.014	00:00:00.005 ± 00:00:00.001
Ling Spam	CHI2	1024	242	83,60 ± 4,44	90,10 ± 10,85	90,21 ± 16,44	69,73 ± 9,40	81,04 ± 8,08	00:00:06.125 ± 00:00:00.016	00:00:00.019 ± 00:00:00.000
Ling Spam	FD	8	4	79,18 ± 10,66	89,45 ± 3,85	88,85 ± 6,29	73,82 ± 14,50	80,49 ± 4,79	00:00:00.497 ± 00:00:00.007	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	16	4	79,19 ± 10,66	89,59 ± 3,87	88,98 ± 6,28	73,82 ± 14,50	80,55 ± 4,80	00:00:00.493 ± 00:00:00.004	00:00:00.002 ± 00:00:00.000
Ling Spam	FD	32	10	92,83 ± 2,69	93,17 ± 1,90	92,35 ± 2,49	93,22 ± 2,74	92,79 ± 0,28	00:00:00.712 ± 00:00:00.020	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	64	19	92,80 ± 3,59	95,47 ± 1,42	95,06 ± 1,71	92,86 ± 3,87	93,90 ± 1,21	00:00:01.107 ± 00:00:00.015	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	128	38	96,06 ± 1,56	96,32 ± 0,53	95,97 ± 0,63	96,32 ± 1,64	96,15 ± 0,57	00:00:02.013 ± 00:00:00.025	00:00:00.005 ± 00:00:00.000
Ling Spam	FD	256	58	97,62 ± 0,65	97,06 ± 0,41	96,76 ± 0,46	97,83 ± 0,63	97,32 ± 0,17	00:00:02.831 ± 00:00:00.024	00:00:00.006 ± 00:00:00.000
Ling Spam	FD	512	103	98,26 ± 0,75	98,38 ± 0,46	98,22 ± 0,53	98,40 ± 0,70	98,31 ± 0,13	00:00:04.482 ± 00:00:00.011	00:00:00.008 ± 00:00:00.000
Ling Spam	FD	1024	121	98,24 ± 0,29	98,53 ± 0,11	98,40 ± 0,11	98,38 ± 0,26	98,39 ± 0,12	00:00:05.376 ± 00:00:00.014	00:00:00.010 ± 00:00:00.000
Ling Spam	MI	8	6	89,09 ± 7,29	90,34 ± 4,50	88,39 ± 6,32	88,08 ± 8,30	88,11 ± 1,30	00:00:00.561 ± 00:00:00.008	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	16	12	91,83 ± 4,30	94,84 ± 2,39	94,17 ± 3,09	91,67 ± 4,54	92,87 ± 0,93	00:00:00.785 ± 00:00:00.017	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	32	19	93,06 ± 2,62	96,58 ± 1,76	96,25 ± 2,18	93,18 ± 2,68	94,65 ± 0,84	00:00:01.097 ± 00:00:00.006	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	64	34	96,20 ± 1,22	97,20 ± 1,27	96,91 ± 1,51	96,45 ± 1,15	96,66 ± 0,21	00:00:01.869 ± 00:00:00.014	00:00:00.004 ± 00:00:00.000
Ling Spam	MI	128	52	97,20 ± 1,23	97,59 ± 0,86	97,35 ± 0,99	97,40 ± 1,16	97,37 ± 0,45	00:00:02.592 ± 00:00:00.016	00:00:00.005 ± 00:00:00.000
Ling Spam	MI	256	68	97,16 ± 1,02	98,06 ± 0,95	97,85 ± 1,12	97,34 ± 0,98	97,59 ± 0,09	00:00:03.242 ± 00:00:00.006	00:00:00.007 ± 00:00:00.001
Ling Spam	MI	512	139	98,31 ± 0,48	98,52 ± 0,13	98,38 ± 0,13	98,46 ± 0,44	98,42 ± 0,21	00:00:06.196 ± 00:00:00.017	00:00:00.011 ± 00:00:00.000
Ling Spam	MI	1024	157	98,18 ± 0,60	98,34 ± 0,52	98,19 ± 0,58	98,32 ± 0,55	98,26 ± 0,09	00:00:06.936 ± 00:00:00.016	00:00:00.012 ± 00:00:00.000
Spam Assassin	CHI2	8	4	68,13 ± 5,41	74,00 ± 13,50	55,77 ± 34,05	69,60 ± 19,94	55,91 ± 12,53	00:00:00.205 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	16	7	69,82 ± 10,01	77,79 ± 10,00	71,93 ± 26,76	59,56 ± 15,85	61,87 ± 9,08	00:00:00.249 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	32	14	72,86 ± 12,33	85,85 ± 8,60	74,82 ± 24,00	61,50 ± 19,89	67,82 ± 8,37	00:00:00.351 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	64	20	72,43 ± 10,64	78,60 ± 11,74	68,27 ± 27,88	64,52 ± 18,35	62,42 ± 7,86	00:00:00.439 ± 00:00:00.003	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	128	25	76,59 ± 8,88	81,92 ± 10,10	75,23 ± 22,00	69,83 ± 16,35	70,53 ± 8,36	00:00:00.529 ± 00:00:00.005	00:00:00.002 ± 00:00:00.000
Spam Assassin	CHI2	256	29	81,15 ± 6,75	83,24 ± 8,71	77,30 ± 18,66	79,01 ± 10,10	77,26 ± 6,62	00:00:00.599 ± 00:00:00.009	00:00:00.002 ± 00:00:00.000
Spam Assassin	CHI2	512	7	72,78 ± 12,31	73,33 ± 15,47	50,59 ± 35,58	65,20 ± 26,01	48,81 ± 7,45	00:00:00.254 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	1024	20	62,06 ± 3,55	83,48 ± 8,37	76,63 ± 25,07	52,61 ± 19,95	63,45 ± 8,36	00:00:00.452 ± 00:00:00.003	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	8	5	69,56 ± 6,96	79,66 ± 8,96	70,79 ± 23,89	64,13 ± 17,85	65,55 ± 7,14	00:00:00.229 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	16	6	87,12 ± 4,47	83,22 ± 5,16	71,42 ± 16,78	88,84 ± 7,97	81,66 ± 4,93	00:00:00.245 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	32	9	79,85 ± 3,68	87,43 ± 1,99	87,50 ± 3,07	78,71 ± 5,86	82,89 ± 2,00	00:00:00.288 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	64	16	84,05 ± 2,09	88,24 ± 1,66	87,67 ± 2,20	84,46 ± 2,71	85,98 ± 0,47	00:00:00.388 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	128	26	90,14 ± 1,39	92,38 ± 0,69	91,89 ± 0,86	90,67 ± 1,45	91,25 ± 0,52	00:00:00.550 ± 00:00:00.006	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	256	53	93,53 ± 0,61	94,94 ± 0,58	94,57 ± 0,64	93,95 ± 0,62	94,25 ± 0,26	00:00:01.018 ± 00:00:00.015	00:00:00.002 ± 00:00:00.000
Spam Assassin	FD	512	67	93,10 ± 1,22	94,07 ± 0,69	93,63 ± 0,80	93,53 ± 1,28	93,58 ± 0,63	00:00:01.264 ± 00:00:00.012	00:00:00.002 ± 00:00:00.000
Spam Assassin	FD	1024	58	83,99 ± 0,88	87,27 ± 0,65	86,57 ± 1,03	84,72 ± 1,28	85,62 ± 0,35	00:00:01.108 ± 00:00:00.028	00:00:00.002 ± 00:00:00.000
Spam Assassin	MI	8	3	52,65 ± 26,19	77,79 ± 12,82	63,10 ± 31,24	78,80 ± 13,40	66,35 ± 15,50	00:00:00.199 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	16	6	85,26 ± 4,35	87,48 ± 1,77	86,24 ± 3,18	85,42 ± 6,23	85,80 ± 2,00	00:00:00.244 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	32	18	86,79 ± 1,35	89,23 ± 2,05	88,14 ± 3,39	87,55 ± 1,92	87,86 ± 0,76	00:00:00.420 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	64	33	88,23 ± 0,69	91,19 ± 0,87	90,66 ± 1,03	88,84 ± 0,83	89,72 ± 0,37	00:00:00.672 ± 00:00:00.013	00:00:00.002 ± 00:00:00.000
Spam Assassin	MI	128	58	94,35 ± 0,44	96,05 ± 0,43	95,78 ± 0,43	94,71 ± 0,40	95,22 ± 0,20	00:00:01.104 ± 00:00:00.029	00:00:00.002 ± 00:00:00.000
Spam Assassin	MI	256	85	93,91 ± 0,67	95,16 ± 0,57	94,80 ± 0,64	94,32 ± 0,68	94,55 ± 0,23	00:00:01.560 ± 00:00:00.004	00:00:00.003 ± 00:00:00.000
Spam Assassin	MI	512	96	93,18 ± 0,66	95,05 ± 0,39	94,73 ± 0,42	93,57 ± 0,65	94,13 ± 0,24	00:00:01.739 ± 00:00:00.008	00:00:00.003 ± 00:00:00.000
Spam Assassin	MI	1024	84	92,57 ± 0,66	94,57 ± 0,54	94,25 ± 0,57	92,98 ± 0,62	93,59 ± 0,34	00:00:01.553 ± 00:00:00.007	00:00:00.003 ± 00:00:00.000
TREC	CHI2	8	4	83,92 ± 14,05	62,50 ± 21,70	40,07 ± 36,89	60,11 ± 36,84	31,81 ± 9,05	00:00:00.041 ± 00:00:00.000	00:00:00.066 ± 00:00:00.001
TREC	CHI2	16	11	79,37 ± 13,57	63,81 ± 18,92	50,19 ± 37,51	50,45 ± 37,35	34,77 ± 9,42	00:00:00.152 ± 00:00:00.008	00:00:00.083 ± 00:00:00.002
TREC	CHI2	32	10	75,57 ± 12,79	65,31 ± 20,01	50,18 ± 37,53	50,36 ± 37,36	34,69 ± 9,41	00:00:00.160 ± 00:00:00.005	00:00:00.081 ± 00:00:00.004
TREC	CHI2	64	12	79,28 ± 13,56	66,65 ± 20,92	50,19 ± 37,53	50,46 ± 37,34	34,76 ± 9,39	00:00:00.217 ± 00:00:00.008	00:00:00.084 ± 00:00:00.006
TREC	CHI2	128	45	78,60 ± 11,15	58,55 ± 18,23	40,43 ± 36,60	60,48 ± 36,35	32,72 ± 9,31	00:00:00.943 ± 00:00:00.022	00:00:00.120 ± 00:00:00.006
TREC	CHI2	256	68	87,57 ± 7,03	44,80 ± 12,39	11,21 ± 22,30	90,02 ± 21,96	26,14 ± 5,64	00:00:02.586 ± 00:00:00.045	00:00:00.156 ± 00:00:00.010
TREC	CHI2	512	133	67,74 ± 9,42	84,01 ± 16,45	80,52 ± 29,13	23,67 ± 28,63	44,91 ± 6,90	00:00:14.459 ± 00:00:00.053	00:00:00.218 ± 00:00:00.006
TREC	CHI2	1024	47	72,84 ± 8,81	56,24 ± 15,31	41,27 ± 35,74	60,86 ± 35,20	34,76 ± 8,41	00:00:15.954 ± 00:00:00.165	00:00:00.097 ± 00:00:00.001
TREC	FD	8	6	57,63 ± 9,63	57,47 ± 10,98	82,70 ± 22,26	24,13 ± 27,46	41,86 ± 7,32	00:00:14.694 ± 00:00:00.257	00:00:00.052 ± 00:00:00.002
TREC	FD	16	7	63,57 ± 11,15	68,62 ± 5,30	77,43 ± 22,64	35,14 ± 29,80	48,71 ± 8,39	00:00:15.404 ± 00:00:00.247	00:00:00.054 ± 00:00:00.003
TREC	FD	32	14	78,98 ± 7,11	73,88 ± 7,99	65,26 ± 21,11	72,84 ± 20,97	68,08 ± 8,67	00:00:20.396 ± 00:00:00.210	00:00:00.053 ± 00:00:00.004
TREC	FD	64	7	72,04 ± 12,62	77,07 ± 14,28	63,22 ± 31,41	59,30 ± 24,50	54,75 ± 5,39	00:00:15.331 ± 00:00:00.107	00:00:00.055 ± 00:00:00.001
TREC	FD	128	22	79,12 ± 6,50	74,01 ± 11,09	58,31 ± 21,37	81,05 ± 14,98	68,10 ± 6,79	00:00:25.042 ± 00:00:00.279	00:00:00.064 ± 00:00:00.004
TREC	FD	256	24	76,07 ± 7,41	83,73 ± 11,43	76,97 ± 21,98	69,76 ± 13,48	71,85 ± 5,04	00:00:26.187 ± 00:00:00.308	00:00:00.065 ± 00:00:00.001
TREC	FD	512	51	73,19 ± 3,90	89,82 ± 7,31	89,04 ± 13,41	66,61 ± 8,33	76,86 ± 4,09	00:00:39.004 ± 00:00:00.507	00:00:00.085 ± 00:00:00.001
TREC	FD	1024	49	74,40 ± 4,85	85,78 ± 8,84	83,16 ± 14,21	69,62 ± 10,59	75,62 ± 2,30	00:00:37.934 ± 00:00:00.095	00:00:00.083 ± 00:00:00.001
TREC	MI	8	5	66,88 ± 12,08	83,67 ± 14,68	75,34 ± 22,75	44,46 ± 30,40	52,62 ± 6,40	00:00:13.683 ± 00:00:00.378	00:00:00.051 ± 00:00:00.002
TREC	MI	16	6	64,50 ± 11,06	88,76 ± 12,52	83,63 ± 19,04	37,80 ± 27,12	54,19 ± 8,76	00:00:14.785 ± 00:00:00.168	00:00:00.052 ± 00:00:00.003
TREC	MI	32	9	73,26 ± 9,17	83,30 ± 10,38	78,71 ± 19,02	64,92 ± 17,64	69,81 ± 5,01	00:00:16.865 ± 00:00:00.256	00:00:00.055 ± 00:00:00.006
TREC	MI	64	16	74,47 ± 7,62	89,50 ± 8,47	85,92 ± 17,39	66,58 ± 12,43	75,21 ± 6,47	00:00:21.659 ± 00:00:00.310	00:00:00.060 ± 00:00:00.007
TREC	MI	128	26	79,83 ± 7,83	84,11 ± 11,90	76,46 ± 22,15	75,26 ± 12,63	74,33 ± 6,68	00:00:28.376 ± 00:00:00.341	00:00:00.066 ± 00:00:00.002
TREC	MI	256	37	74,59 ± 7,40	87,31 ± 9,46	84,14 ± 17,53	66,89 ± 12,08	74,23 ± 3,76	00:00:33.374 ± 00:00:00.548	00:00:00.074 ± 00:00:00.002
TREC	MI	512	66	79,37 ± 6,71	83,64 ± 10,18	78,18 ± 18,71	75,93 ± 10,47	76,05 ± 4,69	00:00:46.900 ± 00:00:00.236	00:00:00.098 ± 00:00:00.006
TREC	MI	1024	204	85,02 ± 4,41	83,93 ± 8,68	79,35 ± 13,70	84,67 ± 6,75	81,67 ± 3,73	00:01:48.535 ± 00:00:00.208	00:00:00.201 ± 00:00:00.000
Unifei 2017	CHI2	8	8	40,95 ± 0,00	89,61 ± 6,09	99,98 ± 0,01	0,09 ± 0,00	23,89 ± 0,00	00:00:00.056 ± 00:00:00.002	00:00:00.322 ± 00:00:00.025
Unifei 2017	CHI2	16	10	65,00 ± 18,36	76,08 ± 13,07	50,06 ± 37,65	50,04 ± 37,67	33,93 ± 7,58	00:00:00.411 ± 00:00:00.005	00:00:00.352 ± 00:00:00.015
Unifei 2017	CHI2	32	20	75,81 ± 17,29	66,62 ± 8,82	30,07 ± 34,50	70,04 ± 34,50	37,98 ± 6,92	00:00:00.751 ± 00:00:00.017	00:00:00.423 ± 00:00:00.018
Unifei 2017	CHI2	64	48	63,95 ± 17,47	74,24 ± 11,49	50,07 ± 37,63	50,05 ± 37,65	33,98 ± 7,62	00:00:01.600 ± 00:00:00.051	00:00:00.553 ± 00:00:00.053
Unifei 2017	CHI2	128	86	58,43 ± 16,57	76,84 ± 10,83	60,11 ± 36,80	40,11 ± 36,84	32,09 ± 7,54	00:00:03.222 ± 00:00:00.043	00:00:00.719 ± 00:00:00.005
Unifei 2017	CHI2	256	34	61,37 ± 19,25	72,63 ± 11,52	59,89 ± 36,80	40,27 ± 36,77	32,13 ± 7,62	00:00:03.865 ± 00:00:00.038	00:00:00.500 ± 00:00:00.049
Unifei 2017	CHI2	512	45	54,28 ± 15,20	67,12 ± 5,18	69,47 ± 33,63	31,16 ± 33,90	31,60 ± 6,68	00:00:18.056 ± 00:00:00.169	00:00:00.536 ± 00:00:00.110
Unifei 2017	CHI2	1024	17	54,90 ± 15,16	90,65 ± 12,66	80,13 ± 29,84	25,45 ± 28,09	35,61 ± 2,45	00:00:50.806 ± 00:00:00.749	00:00:00.286 ± 00:00:00.001
Unifei 2017	FD	8	5	41,28 ± 16,21	64,33 ± 7,10	62,42 ± 31,55	42,51 ± 29,69	42,44 ± 5,41	00:01:08.156 ± 00:00:01.894	00:00:00.205 ± 00:00:00.022
Unifei 2017	FD	16	3	40,21 ± 16,18	65,48 ± 11,99	48,31 ± 36,39	56,78 ± 32,58	39,58 ± 4,70	00:01:02.987 ± 00:00:01.550	00:00:00.209 ± 00:00:00.011
Unifei 2017	FD	32	6	61,98 ± 11,04	67,18 ± 18,64	77,30 ± 28,83	32,62 ± 26,33	45,15 ± 6,82	00:01:09.215 ± 00:00:00.814	00:00:00.223 ± 00:00:00.012
Unifei 2017	FD	64	22	66,29 ± 7,92	75,23 ± 8,61	72,20 ± 24,45	52,96 ± 21,51	59,20 ± 7,21	00:01:51.686 ± 00:00:01.107	00:00:00.269 ± 00:00:00.015
Unifei 2017	FD	128	65	72,44 ± 6,30	84,88 ± 7,93	83,91 ± 18,77	60,53 ± 13,95	70,89 ± 7,51	00:03:16.815 ± 00:00:00.749	00:00:00.403 ± 00:00:00.013
Unifei 2017	FD	256	51	83,91 ± 11,22	69,82 ± 10,24	53,63 ± 23,63	79,72 ± 17,82	63,45 ± 4,33	00:02:47.959 ± 00:00:00.836	00:00:00.359 ± 00:00:00.017
Unifei 2017	FD	512	101	82,00 ± 9,82	80,25 ± 9,41	73,00 ± 17,53	77,56 ± 16,95	74,13 ± 7,10	00:04:21.217 ± 00:00:00.824	00:00:00.494 ± 00:00:00.017
Unifei 2017	FD	1024	105	87,84 ± 6,27	77,55 ± 8,13	67,51 ± 16,19	87,91 ± 11,67	77,32 ± 6,45	00:04:28.736 ± 00:00:01.038	00:00:00.501 ± 00:00:00.011
Unifei 2017	MI	8	4	36,64 ± 23,13	58,30 ± 6,19	18,35 ± 24,86	86,39 ± 20,62	43,60 ± 7,16	00:00:54.730 ± 00:00:01.674	00:00:00.211 ± 00:00:00.022
Unifei 2017	MI	16	7	52,18 ± 13,91	59,70 ± 10,17	28,30 ± 32,10	77,50 ± 25,70	40,51 ± 6,79	00:01:06.600 ± 00:00:00.934	00:00:00.226 ± 00:00:00.010
Unifei 2017	MI	32	5	62,42 ± 6,59	68,20 ± 13,09	48,70 ± 36,40	57,76 ± 31,97	41,23 ± 5,52	00:01:02.644 ± 00:00:01.398	00:00:00.213 ± 00:00:00.009
Unifei 2017	MI	64	9	76,41 ± 14,32	74,19 ± 15,27	56,74 ± 31,31	63,39 ± 26,93	52,93 ± 3,83	00:01:16.465 ± 00:00:00.613	00:00:00.243 ± 00:00:00.014
Unifei 2017	MI	128	18	71,85 ± 12,51	78,25 ± 12,06	71,15 ± 27,26	55,92 ± 22,15	58,86 ± 5,61	00:01:39.572 ± 00:00:00.832	00:00:00.258 ± 00:00:00.008
Unifei 2017	MI	256	31	81,94 ± 10,51	72,34 ± 11,37	58,60 ± 22,75	75,69 ± 20,98	64,50 ± 3,51	00:02:12.608 ± 00:00:00.957	00:00:00.299 ± 00:00:00.009
Unifei 2017	MI	512	125	89,05 ± 7,97	75,03 ± 10,32	62,20 ± 18,82	86,59 ± 13,82	73,34 ± 5,61	00:05:11.194 ± 00:00:00.911	00:00:00.587 ± 00:00:00.020
Unifei 2017	MI	1024	208	79,20 ± 5,96	84,68 ± 7,06	83,09 ± 11,72	75,23 ± 9,55	78,78 ± 1,32	00:07:43.886 ± 00:00:00.809	00:00:00.839 ± 00:00:00.011
Unifei 2018	CHI2	8	4	99,79 ± 0,03	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	44,61 ± 0,00	00:00:00.390 ± 00:00:00.017	00:00:00.270 ± 00:00:00.008
Unifei 2018	CHI2	16	16	99,56 ± 0,08	59,17 ± 0,00	2,04 ± 0,01	99,99 ± 0,00	45,27 ± 0,00	00:00:01.705 ± 00:00:00.042	00:00:00.391 ± 00:00:00.036
Unifei 2018	CHI2	32	28	99,36 ± 0,14	63,47 ± 9,09	12,39 ± 21,93	99,90 ± 0,21	51,34 ± 12,07	00:00:04.145 ± 00:00:00.114	00:00:00.427 ± 00:00:00.045
Unifei 2018	CHI2	64	42	57,89 ± 18,77	90,79 ± 11,72	80,37 ± 29,26	31,22 ± 33,20	37,77 ± 15,95	00:00:07.312 ± 00:00:00.046	00:00:00.475 ± 00:00:00.061
Unifei 2018	CHI2	128	69	87,13 ± 16,55	70,28 ± 11,97	35,42 ± 31,11	80,61 ± 28,73	50,63 ± 12,98	00:00:26.123 ± 00:00:00.124	00:00:00.587 ± 00:00:00.037
Unifei 2018	CHI2	256	116	77,42 ± 18,84	74,43 ± 13,03	44,55 ± 33,81	64,55 ± 32,96	45,48 ± 7,43	00:00:40.276 ± 00:00:00.162	00:00:00.825 ± 00:00:00.037
Unifei 2018	CHI2	512	121	64,15 ± 15,98	83,30 ± 12,77	74,02 ± 28,85	39,54 ± 29,66	46,22 ± 3,21	00:02:08.016 ± 00:00:00.358	00:00:00.844 ± 00:00:00.024
Unifei 2018	CHI2	1024	152	72,98 ± 14,89	77,51 ± 13,53	63,07 ± 31,57	58,72 ± 29,13	53,29 ± 9,58	00:03:57.466 ± 00:00:00.587	00:00:00.856 ± 00:00:00.020
Unifei 2018	FD	8	5	50,95 ± 13,30	59,14 ± 6,58	40,76 ± 34,05	62,20 ± 32,72	40,08 ± 2,11	00:01:06.461 ± 00:00:00.316	00:00:00.208 ± 00:00:00.017
Unifei 2018	FD	16	3	42,10 ± 15,87	76,62 ± 10,50	77,61 ± 29,28	31,10 ± 26,01	43,41 ± 3,74	00:01:02.120 ± 00:00:01.600	00:00:00.203 ± 00:00:00.004
Unifei 2018	FD	32	8	70,28 ± 11,44	58,75 ± 8,49	30,68 ± 31,92	73,39 ± 29,63	39,39 ± 4,66	00:01:13.114 ± 00:00:00.984	00:00:00.234 ± 00:00:00.005
Unifei 2018	FD	64	11	66,20 ± 10,80	69,79 ± 11,88	57,45 ± 34,95	51,42 ± 30,20	43,97 ± 6,23	00:01:21.125 ± 00:00:00.658	00:00:00.238 ± 00:00:00.006
Unifei 2018	FD	128	24	67,61 ± 8,24	83,64 ± 8,57	85,13 ± 20,22	45,71 ± 19,04	62,26 ± 9,36	00:01:55.815 ± 00:00:00.633	00:00:00.263 ± 00:00:00.008
Unifei 2018	FD	256	60	77,65 ± 8,73	82,43 ± 9,14	76,03 ± 18,42	71,87 ± 16,11	73,66 ± 6,64	00:03:04.783 ± 00:00:00.616	00:00:00.358 ± 00:00:00.015
Unifei 2018	FD	512	79	73,55 ± 6,74	85,10 ± 3,27	86,36 ± 7,46	65,77 ± 11,78	75,71 ± 3,59	00:03:41.589 ± 00:00:00.872	00:00:00.426 ± 00:00:00.004
Unifei 2018	FD	1024	77	75,40 ± 10,26	79,62 ± 11,27	71,80 ± 22,75	66,34 ± 18,74	66,76 ± 4,14	00:03:39.034 ± 00:00:00.686	00:00:00.427 ± 00:00:00.014
Unifei 2018	MI	8	3	63,68 ± 11,53	38,04 ± 18,89	42,26 ± 30,17	63,16 ± 31,68	42,92 ± 6,99	00:01:19.300 ± 00:00:01.258	00:00:00.282 ± 00:00:00.015
Unifei 2018	MI	16	3	54,37 ± 14,55	69,68 ± 5,98	70,95 ± 21,59	48,74 ± 21,00	56,17 ± 6,47	00:01:00.769 ± 00:00:01.818	00:00:00.203 ± 00:00:00.009
Unifei 2018	MI	32	7	57,74 ± 24,31	65,68 ± 7,86	46,25 ± 29,81	71,90 ± 22,29	53,21 ± 9,40	00:01:09.481 ± 00:00:00.707	00:00:00.229 ± 00:00:00.011
Unifei 2018	MI	64	10	80,60 ± 14,00	72,65 ± 14,43	52,94 ± 31,05	71,19 ± 25,87	55,37 ± 10,00	00:01:18.644 ± 00:00:00.447	00:00:00.229 ± 00:00:00.016
Unifei 2018	MI	128	30	66,57 ± 10,22	83,49 ± 10,75	80,47 ± 23,36	48,12 ± 19,80	59,87 ± 4,45	00:02:09.671 ± 00:00:02.518	00:00:00.294 ± 00:00:00.014
Unifei 2018	MI	256	49	74,84 ± 11,40	80,93 ± 10,28	74,31 ± 22,40	64,15 ± 20,52	66,75 ± 6,38	00:02:42.530 ± 00:00:00.952	00:00:00.337 ± 00:00:00.008
Unifei 2018	MI	512	118	81,65 ± 7,34	83,83 ± 7,01	80,36 ± 11,94	78,69 ± 13,31	79,17 ± 4,57	00:04:52.090 ± 00:00:00.531	00:00:00.540 ± 00:00:00.012
Unifei 2018	MI	1024	155	78,33 ± 6,65	84,68 ± 7,37	82,56 ± 12,56	74,22 ± 10,34	77,86 ± 1,19	00:06:02.319 ± 00:00:00.830	00:00:00.684 ± 00:00:00.021

Tabela 21: Resultados do classificador RBF.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.012 ± 00:00:00.008	00:00:00.019 ± 00:00:00.005
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	77,42 ± 0,04	00:00:00.017 ± 00:00:00.016	00:00:00.017 ± 00:00:00.003
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	78,40 ± 0,03	00:00:00.035 ± 00:00:00.043	00:00:00.015 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,19 ± 0,17	99,93 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.033 ± 00:00:00.010	00:00:00.029 ± 00:00:00.002
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,41 ± 0,09	99,91 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.241 ± 00:00:00.192	00:00:00.037 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,74 ± 0,08	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:01.149 ± 00:00:01.786	00:00:00.178 ± 00:00:00.004
Ling Spam	CHI2	512	14	68,54 ± 0,18	90,57 ± 0,50	97,25 ± 0,19	37,10 ± 0,65	68,88 ± 0,29	00:00:01.061 ± 00:00:00.194	00:00:00.253 ± 00:00:00.005
Ling Spam	CHI2	1024	242	73,12 ± 2,42	94,91 ± 4,29	97,81 ± 2,31	45,85 ± 7,08	74,46 ± 2,34	00:00:14.224 ± 00:00:00.692	00:00:11.594 ± 00:00:00.019
Ling Spam	FD	8	4	97,40 ± 0,23	79,41 ± 2,16	72,04 ± 3,59	98,23 ± 0,22	85,37 ± 1,76	00:00:00.780 ± 00:00:00.106	00:00:00.044 ± 00:00:00.001
Ling Spam	FD	16	4	97,20 ± 0,34	83,04 ± 0,84	78,15 ± 1,26	97,94 ± 0,25	88,32 ± 0,63	00:00:01.791 ± 00:00:00.403	00:00:00.068 ± 00:00:00.000
Ling Spam	FD	32	10	96,58 ± 1,67	88,83 ± 1,30	86,57 ± 1,87	97,08 ± 1,67	92,03 ± 0,39	00:00:02.445 ± 00:00:00.907	00:00:00.173 ± 00:00:00.001
Ling Spam	FD	64	19	98,90 ± 0,28	90,05 ± 0,50	88,05 ± 0,64	99,10 ± 0,23	93,78 ± 0,27	00:00:02.864 ± 00:00:00.188	00:00:00.392 ± 00:00:00.001
Ling Spam	FD	128	38	99,02 ± 0,55	89,59 ± 0,43	87,41 ± 0,56	99,20 ± 0,47	93,53 ± 0,33	00:00:04.814 ± 00:00:00.286	00:00:00.966 ± 00:00:00.003
Ling Spam	FD	256	58	99,55 ± 0,25	91,83 ± 0,34	90,32 ± 0,44	99,63 ± 0,21	95,16 ± 0,20	00:00:08.640 ± 00:00:01.003	00:00:01.817 ± 00:00:00.005
Ling Spam	FD	512	103	99,86 ± 0,15	92,31 ± 0,21	90,92 ± 0,26	99,89 ± 0,12	95,58 ± 0,11	00:00:14.850 ± 00:00:01.326	00:00:04.007 ± 00:00:00.018
Ling Spam	FD	1024	121	99,71 ± 0,19	89,65 ± 0,39	87,43 ± 0,50	99,77 ± 0,16	93,83 ± 0,24	00:00:19.898 ± 00:00:01.602	00:00:05.705 ± 00:00:00.014
Ling Spam	MI	8	6	73,32 ± 1,53	96,48 ± 0,42	97,31 ± 0,33	67,45 ± 2,56	81,39 ± 1,37	00:00:00.726 ± 00:00:00.091	00:00:00.051 ± 00:00:00.000
Ling Spam	MI	16	12	92,81 ± 3,77	91,40 ± 2,72	90,03 ± 3,62	93,08 ± 4,04	91,60 ± 0,53	00:00:01.760 ± 00:00:00.769	00:00:00.130 ± 00:00:00.000
Ling Spam	MI	32	19	90,88 ± 2,71	95,26 ± 2,41	94,78 ± 2,98	90,98 ± 2,97	92,79 ± 0,37	00:00:02.190 ± 00:00:00.412	00:00:00.277 ± 00:00:00.001
Ling Spam	MI	64	34	95,24 ± 0,71	96,89 ± 1,18	96,60 ± 1,45	95,55 ± 0,71	96,06 ± 0,39	00:00:03.622 ± 00:00:00.232	00:00:00.641 ± 00:00:00.001
Ling Spam	MI	128	52	97,24 ± 1,37	94,78 ± 1,34	94,06 ± 1,66	97,49 ± 1,29	95,84 ± 0,22	00:00:06.193 ± 00:00:00.674	00:00:01.258 ± 00:00:00.002
Ling Spam	MI	256	68	99,19 ± 0,18	93,10 ± 0,37	91,97 ± 0,44	99,31 ± 0,15	95,79 ± 0,20	00:00:09.002 ± 00:00:00.826	00:00:02.094 ± 00:00:00.005
Ling Spam	MI	512	139	99,02 ± 0,55	95,35 ± 0,94	94,70 ± 1,12	99,13 ± 0,50	97,01 ± 0,32	00:00:18.229 ± 00:00:00.918	00:00:05.242 ± 00:00:00.021
Ling Spam	MI	1024	157	98,66 ± 0,32	93,73 ± 0,40	92,78 ± 0,51	98,84 ± 0,29	95,93 ± 0,19	00:00:23.963 ± 00:00:01.526	00:00:07.263 ± 00:00:00.020
Spam Assassin	CHI2	8	4	66,84 ± 0,90	76,80 ± 1,15	78,40 ± 1,80	64,65 ± 2,06	71,09 ± 0,46	00:00:00.141 ± 00:00:00.007	00:00:00.019 ± 00:00:00.000
Spam Assassin	CHI2	16	7	66,93 ± 1,14	79,37 ± 0,86	81,76 ± 1,26	63,37 ± 2,38	71,93 ± 0,93	00:00:00.327 ± 00:00:00.024	00:00:00.041 ± 00:00:00.000
Spam Assassin	CHI2	32	14	64,82 ± 2,31	81,39 ± 1,03	85,38 ± 1,93	57,62 ± 5,30	70,25 ± 2,37	00:00:00.524 ± 00:00:00.038	00:00:00.096 ± 00:00:00.001
Spam Assassin	CHI2	64	20	63,33 ± 1,76	83,11 ± 1,91	87,48 ± 1,93	54,43 ± 3,99	69,29 ± 1,82	00:00:01.264 ± 00:00:00.923	00:00:00.179 ± 00:00:00.001
Spam Assassin	CHI2	128	25	63,51 ± 2,67	84,79 ± 3,24	88,93 ± 3,65	53,06 ± 6,46	69,02 ± 2,44	00:00:01.019 ± 00:00:00.036	00:00:00.287 ± 00:00:00.001
Spam Assassin	CHI2	256	29	72,93 ± 2,10	85,94 ± 1,86	87,21 ± 2,30	70,30 ± 3,41	78,21 ± 1,21	00:00:01.571 ± 00:00:00.058	00:00:00.416 ± 00:00:00.001
Spam Assassin	CHI2	512	7	62,51 ± 0,33	92,96 ± 0,46	96,05 ± 0,26	47,51 ± 0,28	69,01 ± 0,21	00:00:01.736 ± 00:00:00.370	00:00:00.187 ± 00:00:00.004
Spam Assassin	CHI2	1024	20	67,90 ± 0,81	85,05 ± 1,94	88,29 ± 1,80	61,32 ± 1,71	73,87 ± 0,72	00:00:02.258 ± 00:00:00.141	00:00:00.481 ± 00:00:00.006
Spam Assassin	FD	8	5	65,36 ± 2,14	81,63 ± 2,99	85,22 ± 4,12	58,10 ± 5,32	70,46 ± 1,58	00:00:00.188 ± 00:00:00.014	00:00:00.020 ± 00:00:00.000
Spam Assassin	FD	16	6	79,13 ± 1,03	86,19 ± 0,98	86,27 ± 1,18	78,95 ± 1,49	82,46 ± 0,64	00:00:00.391 ± 00:00:00.037	00:00:00.036 ± 00:00:00.000
Spam Assassin	FD	32	9	82,97 ± 1,74	87,02 ± 1,32	86,43 ± 1,79	83,51 ± 2,24	84,90 ± 0,61	00:00:00.626 ± 00:00:00.040	00:00:00.070 ± 00:00:00.000
Spam Assassin	FD	64	16	86,13 ± 0,58	91,07 ± 0,84	90,80 ± 0,91	86,50 ± 0,62	88,57 ± 0,48	00:00:01.043 ± 00:00:00.080	00:00:00.148 ± 00:00:00.000
Spam Assassin	FD	128	26	91,87 ± 0,77	91,01 ± 1,02	90,05 ± 1,33	92,61 ± 0,87	91,38 ± 0,41	00:00:01.766 ± 00:00:00.207	00:00:00.295 ± 00:00:00.000
Spam Assassin	FD	256	53	93,47 ± 1,11	93,36 ± 1,04	92,74 ± 1,23	93,98 ± 1,13	93,38 ± 0,47	00:00:02.782 ± 00:00:00.221	00:00:00.729 ± 00:00:00.002
Spam Assassin	FD	512	67	94,51 ± 1,33	91,27 ± 0,91	90,13 ± 1,11	95,12 ± 1,31	92,72 ± 0,72	00:00:03.893 ± 00:00:00.262	00:00:01.132 ± 00:00:00.004
Spam Assassin	FD	1024	58	78,13 ± 2,08	88,24 ± 1,01	88,83 ± 1,43	76,86 ± 3,08	82,56 ± 1,16	00:00:04.482 ± 00:00:00.131	00:00:01.261 ± 00:00:00.004
Spam Assassin	MI	8	3	78,85 ± 0,52	86,42 ± 0,68	86,48 ± 0,90	78,74 ± 0,77	82,44 ± 0,29	00:00:00.243 ± 00:00:00.019	00:00:00.019 ± 00:00:00.003
Spam Assassin	MI	16	6	83,75 ± 2,78	89,54 ± 1,61	89,14 ± 2,28	83,80 ± 3,76	86,35 ± 1,17	00:00:00.462 ± 00:00:00.091	00:00:00.037 ± 00:00:00.000
Spam Assassin	MI	32	18	83,77 ± 1,31	94,67 ± 0,96	94,94 ± 0,93	82,97 ± 1,66	88,70 ± 0,95	00:00:00.786 ± 00:00:00.061	00:00:00.115 ± 00:00:00.002
Spam Assassin	MI	64	33	84,19 ± 1,01	96,23 ± 1,09	96,45 ± 1,07	83,25 ± 1,29	89,56 ± 0,58	00:00:01.301 ± 00:00:00.069	00:00:00.271 ± 00:00:00.002
Spam Assassin	MI	128	58	89,00 ± 0,99	97,44 ± 0,35	97,46 ± 0,37	88,87 ± 1,08	92,99 ± 0,50	00:00:02.530 ± 00:00:00.242	00:00:00.602 ± 00:00:00.001
Spam Assassin	MI	256	85	87,38 ± 0,81	97,03 ± 0,61	97,10 ± 0,64	87,03 ± 0,99	91,86 ± 0,50	00:00:03.907 ± 00:00:00.281	00:00:01.124 ± 00:00:00.004
Spam Assassin	MI	512	96	88,00 ± 1,07	95,77 ± 0,77	95,78 ± 0,85	87,91 ± 1,30	91,68 ± 0,50	00:00:04.839 ± 00:00:00.289	00:00:01.592 ± 00:00:00.005
Spam Assassin	MI	1024	84	93,84 ± 1,37	92,29 ± 0,94	91,39 ± 1,33	94,40 ± 1,37	92,95 ± 0,35	00:00:05.418 ± 00:00:00.294	00:00:01.719 ± 00:00:00.004
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	46,53 ± 0,01	00:00:00.013 ± 00:00:00.001	00:00:00.550 ± 00:00:00.016
TREC	CHI2	16	11	61,42 ± 0,00	90,70 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	47,31 ± 0,01	00:00:00.125 ± 00:00:00.021	00:00:01.864 ± 00:00:00.061
TREC	CHI2	32	10	61,39 ± 0,00	90,24 ± 0,73	99,94 ± 0,00	0,89 ± 0,02	47,23 ± 0,01	00:00:00.117 ± 00:00:00.023	00:00:01.564 ± 00:00:00.069
TREC	CHI2	64	12	61,37 ± 0,00	93,56 ± 0,39	99,96 ± 0,00	1,02 ± 0,01	47,29 ± 0,01	00:00:00.144 ± 00:00:00.013	00:00:02.525 ± 00:00:00.133
TREC	CHI2	128	45	61,70 ± 0,01	87,92 ± 0,25	99,84 ± 0,00	1,82 ± 0,02	48,13 ± 0,02	00:00:01.324 ± 00:00:00.064	00:00:22.208 ± 00:00:00.078
TREC	CHI2	256	68	86,52 ± 0,39	39,39 ± 0,00	1,91 ± 0,02	99,54 ± 0,01	24,31 ± 0,02	00:00:03.598 ± 00:00:00.288	00:00:34.862 ± 00:00:00.460
TREC	CHI2	512	133	64,64 ± 6,99	88,37 ± 12,07	90,13 ± 21,81	14,62 ± 21,36	47,52 ± 5,23	00:00:28.705 ± 00:00:02.306	00:01:34.099 ± 00:00:00.184
TREC	CHI2	1024	47	61,67 ± 0,10	77,42 ± 1,52	98,67 ± 0,18	6,87 ± 0,54	50,76 ± 0,38	00:01:05.142 ± 00:00:13.949	00:00:40.039 ± 00:00:00.343
TREC	FD	8	6	63,06 ± 2,51	72,11 ± 2,26	77,39 ± 6,39	54,99 ± 7,35	65,32 ± 1,34	00:00:08.571 ± 00:00:00.446	00:00:00.695 ± 00:00:00.003
TREC	FD	16	7	65,89 ± 3,49	72,95 ± 1,74	76,30 ± 5,23	60,65 ± 7,15	67,95 ± 1,50	00:00:15.363 ± 00:00:01.054	00:00:01.227 ± 00:00:00.010
TREC	FD	32	14	75,17 ± 2,68	71,71 ± 2,42	67,72 ± 5,54	77,74 ± 5,01	72,60 ± 1,06	00:00:34.066 ± 00:00:02.386	00:00:02.917 ± 00:00:00.016
TREC	FD	64	7	66,04 ± 0,20	86,40 ± 0,70	91,10 ± 0,53	54,67 ± 0,34	71,69 ± 0,29	00:00:44.143 ± 00:00:08.637	00:00:02.563 ± 00:00:00.017
TREC	FD	128	22	71,22 ± 0,48	84,80 ± 0,96	87,78 ± 1,06	65,71 ± 1,02	76,29 ± 0,40	00:01:44.350 ± 00:00:10.950	00:00:07.970 ± 00:00:00.034
TREC	FD	256	24	69,17 ± 0,78	86,80 ± 0,51	90,35 ± 0,44	61,15 ± 1,56	74,98 ± 0,82	00:02:16.447 ± 00:00:35.922	00:00:11.013 ± 00:00:00.051
TREC	FD	512	51	70,25 ± 1,81	85,77 ± 1,22	88,97 ± 1,56	63,52 ± 3,58	75,60 ± 1,48	00:04:03.796 ± 00:00:30.361	00:00:27.419 ± 00:00:00.073
TREC	FD	1024	49	67,88 ± 1,24	84,37 ± 1,43	88,38 ± 1,63	59,72 ± 2,77	73,22 ± 1,06	00:04:34.523 ± 00:01:06.118	00:00:32.509 ± 00:00:00.182
TREC	MI	8	5	80,92 ± 1,82	61,15 ± 0,99	45,12 ± 3,11	88,87 ± 1,93	64,98 ± 1,17	00:00:08.375 ± 00:00:00.870	00:00:00.658 ± 00:00:00.007
TREC	MI	16	6	79,45 ± 1,17	64,75 ± 0,72	52,68 ± 2,38	86,34 ± 1,49	68,63 ± 0,69	00:00:16.768 ± 00:00:01.326	00:00:01.129 ± 00:00:00.009
TREC	MI	32	9	63,83 ± 0,28	90,71 ± 1,00	94,88 ± 0,68	48,05 ± 0,82	69,44 ± 0,32	00:00:28.062 ± 00:00:02.457	00:00:02.125 ± 00:00:00.010
TREC	MI	64	16	66,83 ± 1,06	92,18 ± 0,80	95,21 ± 0,75	54,14 ± 2,52	73,24 ± 1,16	00:00:51.312 ± 00:00:07.368	00:00:04.525 ± 00:00:00.016
TREC	MI	128	26	71,05 ± 1,70	90,05 ± 1,46	92,66 ± 1,57	63,21 ± 3,63	77,20 ± 1,42	00:01:54.845 ± 00:00:24.283	00:00:08.976 ± 00:00:00.026
TREC	MI	256	37	69,44 ± 4,48	83,41 ± 8,65	82,18 ± 16,36	61,97 ± 11,35	70,58 ± 3,16	00:01:54.256 ± 00:00:26.719	00:00:16.053 ± 00:00:00.141
TREC	MI	512	66	77,42 ± 7,01	71,08 ± 8,19	61,28 ± 16,76	78,83 ± 12,04	68,74 ± 3,56	00:02:54.129 ± 00:00:18.504	00:00:34.254 ± 00:00:00.215
TREC	MI	1024	204	70,33 ± 2,23	79,40 ± 4,34	81,08 ± 6,75	67,09 ± 3,81	73,73 ± 2,79	00:12:05.186 ± 00:00:52.664	00:02:04.511 ± 00:00:00.347
Unifei 2017	CHI2	8	8	64,61 ± 17,84	78,11 ± 14,34	50,06 ± 37,66	50,04 ± 37,67	33,93 ± 7,57	00:00:00.054 ± 00:00:00.026	00:00:05.118 ± 00:00:00.069
Unifei 2017	CHI2	16	10	40,96 ± 0,00	88,57 ± 0,70	99,98 ± 0,00	0,13 ± 0,00	23,94 ± 0,00	00:00:00.285 ± 00:00:00.009	00:00:07.681 ± 00:00:00.213
Unifei 2017	CHI2	32	20	40,95 ± 0,00	84,99 ± 0,75	99,96 ± 0,00	0,17 ± 0,00	23,98 ± 0,00	00:00:00.698 ± 00:00:00.055	00:00:23.656 ± 00:00:00.483
Unifei 2017	CHI2	64	48	40,92 ± 0,00	84,02 ± 0,74	99,94 ± 0,00	0,21 ± 0,00	23,98 ± 0,00	00:00:01.735 ± 00:00:00.235	00:01:14.049 ± 00:00:00.548
Unifei 2017	CHI2	128	86	40,85 ± 0,00	85,35 ± 0,38	99,93 ± 0,00	0,28 ± 0,00	24,00 ± 0,00	00:00:03.845 ± 00:00:00.151	00:02:49.926 ± 00:00:00.453
Unifei 2017	CHI2	256	34	40,56 ± 0,00	68,38 ± 0,36	99,46 ± 0,01	0,79 ± 0,01	24,26 ± 0,01	00:00:05.174 ± 00:00:00.482	00:01:15.810 ± 00:00:01.560
Unifei 2017	CHI2	512	45	41,34 ± 0,01	64,69 ± 0,11	96,86 ± 0,01	4,02 ± 0,01	28,29 ± 0,01	00:01:02.909 ± 00:00:07.020	00:02:26.025 ± 00:00:02.335
Unifei 2017	CHI2	1024	17	44,97 ± 0,02	98,47 ± 0,62	99,85 ± 0,06	7,27 ± 0,03	34,45 ± 0,03	00:01:51.722 ± 00:00:20.385	00:00:53.773 ± 00:00:00.786
Unifei 2017	FD	8	5	60,57 ± 1,28	61,38 ± 0,94	62,63 ± 2,98	59,14 ± 3,68	60,80 ± 0,95	00:00:30.867 ± 00:00:00.660	00:00:02.530 ± 00:00:00.014
Unifei 2017	FD	16	3	59,53 ± 0,45	72,52 ± 1,18	83,42 ± 1,73	43,37 ± 2,17	61,81 ± 0,63	00:00:56.147 ± 00:00:01.886	00:00:03.312 ± 00:00:00.014
Unifei 2017	FD	32	6	62,67 ± 0,56	72,12 ± 0,76	79,52 ± 1,47	52,70 ± 2,00	65,45 ± 0,56	00:01:48.314 ± 00:00:07.276	00:00:06.700 ± 00:00:00.023
Unifei 2017	FD	64	22	68,37 ± 0,36	77,47 ± 1,47	81,74 ± 1,83	62,28 ± 1,17	71,72 ± 0,46	00:03:39.700 ± 00:00:14.159	00:00:22.885 ± 00:00:00.135
Unifei 2017	FD	128	65	72,40 ± 0,87	85,99 ± 1,62	89,11 ± 1,67	66,08 ± 1,82	77,27 ± 0,72	00:08:49.833 ± 00:00:26.864	00:01:21.787 ± 00:00:00.153
Unifei 2017	FD	256	51	70,66 ± 1,37	80,58 ± 3,34	83,59 ± 5,40	65,09 ± 3,99	74,00 ± 0,77	00:11:42.405 ± 00:01:42.422	00:01:27.061 ± 00:00:00.170
Unifei 2017	FD	512	101	72,11 ± 1,36	80,54 ± 3,61	82,82 ± 5,96	67,79 ± 3,93	75,05 ± 1,38	00:19:17.764 ± 00:02:47.713	00:03:27.195 ± 00:00:00.635
Unifei 2017	FD	1024	105	70,60 ± 2,88	78,95 ± 5,61	80,39 ± 9,83	65,42 ± 7,48	72,37 ± 1,43	00:26:59.378 ± 00:07:21.559	00:04:26.630 ± 00:00:00.447
Unifei 2017	MI	8	4	61,19 ± 0,22	66,91 ± 0,40	69,52 ± 0,93	58,28 ± 0,77	63,65 ± 0,21	00:00:26.519 ± 00:00:01.349	00:00:02.550 ± 00:00:00.026
Unifei 2017	MI	16	7	57,14 ± 0,21	78,52 ± 1,12	89,85 ± 0,74	35,46 ± 0,53	59,12 ± 0,36	00:00:52.093 ± 00:00:03.328	00:00:04.936 ± 00:00:00.058
Unifei 2017	MI	32	5	57,71 ± 0,23	80,55 ± 1,30	90,80 ± 0,86	36,34 ± 0,70	60,09 ± 0,39	00:01:34.974 ± 00:00:08.426	00:00:06.181 ± 00:00:00.028
Unifei 2017	MI	64	9	62,92 ± 0,22	81,44 ± 0,74	89,14 ± 0,70	47,50 ± 0,82	66,88 ± 0,25	00:04:13.306 ± 00:00:34.562	00:00:12.039 ± 00:00:00.056
Unifei 2017	MI	128	18	65,00 ± 0,49	82,89 ± 0,98	89,23 ± 0,90	51,98 ± 1,33	69,53 ± 0,59	00:07:22.218 ± 00:01:41.808	00:00:26.065 ± 00:00:00.106
Unifei 2017	MI	256	31	70,98 ± 2,85	77,68 ± 4,89	79,05 ± 9,24	66,67 ± 7,01	72,42 ± 1,69	00:08:40.368 ± 00:02:36.400	00:00:53.116 ± 00:00:00.123
Unifei 2017	MI	512	125	74,65 ± 1,00	81,71 ± 1,96	83,73 ± 2,60	71,53 ± 2,36	77,52 ± 0,23	00:24:12.225 ± 00:03:07.822	00:04:07.769 ± 00:00:00.561
Unifei 2017	MI	1024	208	73,89 ± 1,05	79,67 ± 3,38	81,29 ± 4,67	71,43 ± 1,52	76,26 ± 1,96	00:41:28.294 ± 00:05:31.458	00:08:19.482 ± 00:00:00.651
Unifei 2018	CHI2	8	4	91,41 ± 9,77	59,04 ± 0,03	1,12 ± 0,10	99,90 ± 0,11	44,54 ± 0,09	00:00:00.312 ± 00:00:00.078	00:00:03.741 ± 00:00:00.123
Unifei 2018	CHI2	16	16	41,65 ± 0,00	99,40 ± 0,17	99,99 ± 0,00	1,32 ± 0,01	25,83 ± 0,00	00:00:01.806 ± 00:00:00.863	00:00:13.597 ± 00:00:00.100
Unifei 2018	CHI2	32	28	53,26 ± 17,49	91,40 ± 12,04	80,53 ± 29,34	21,51 ± 29,59	30,33 ± 5,91	00:00:04.706 ± 00:00:00.872	00:00:31.525 ± 00:00:00.111
Unifei 2018	CHI2	64	42	41,64 ± 0,01	98,56 ± 0,14	99,95 ± 0,01	2,56 ± 0,01	27,06 ± 0,02	00:00:08.270 ± 00:00:01.918	00:01:03.020 ± 00:00:00.178
Unifei 2018	CHI2	128	69	98,44 ± 0,27	59,89 ± 0,01	8,46 ± 0,02	99,90 ± 0,02	49,85 ± 0,01	00:00:39.648 ± 00:00:04.610	00:02:13.850 ± 00:00:00.386
Unifei 2018	CHI2	256	116	97,08 ± 0,24	60,45 ± 0,03	8,50 ± 0,10	99,82 ± 0,02	50,44 ± 0,08	00:01:17.199 ± 00:00:09.777	00:04:39.345 ± 00:00:00.294
Unifei 2018	CHI2	512	121	55,34 ± 10,14	90,95 ± 8,26	90,90 ± 18,20	24,50 ± 18,79	46,86 ± 1,68	00:05:33.273 ± 00:00:29.103	00:05:48.976 ± 00:00:00.901
Unifei 2018	CHI2	1024	152	60,84 ± 10,58	83,93 ± 9,95	82,14 ± 22,86	38,38 ± 22,30	53,00 ± 0,53	00:16:40.209 ± 00:02:43.068	00:07:43.768 ± 00:00:00.877
Unifei 2018	FD	8	5	62,81 ± 0,94	68,42 ± 1,07	73,02 ± 1,49	57,45 ± 1,97	64,94 ± 0,93	00:00:32.472 ± 00:00:00.924	00:00:02.573 ± 00:00:00.029
Unifei 2018	FD	16	3	60,15 ± 0,30	78,06 ± 1,38	87,56 ± 1,41	43,10 ± 1,50	63,31 ± 0,41	00:00:54.363 ± 00:00:01.497	00:00:03.295 ± 00:00:00.011
Unifei 2018	FD	32	8	62,27 ± 1,10	72,50 ± 1,39	79,26 ± 2,01	53,10 ± 2,82	65,40 ± 1,19	00:01:50.542 ± 00:00:06.781	00:00:07.904 ± 00:00:00.033
Unifei 2018	FD	64	11	62,17 ± 1,08	80,49 ± 1,76	87,97 ± 2,01	47,73 ± 3,28	66,19 ± 1,27	00:03:28.128 ± 00:00:08.479	00:00:13.503 ± 00:00:00.044
Unifei 2018	FD	128	24	69,22 ± 1,11	81,22 ± 1,27	84,94 ± 1,66	63,11 ± 2,34	73,57 ± 0,88	00:06:21.575 ± 00:00:23.333	00:00:33.276 ± 00:00:00.068
Unifei 2018	FD	256	60	70,88 ± 0,90	81,73 ± 2,98	84,43 ± 4,01	66,16 ± 2,42	74,94 ± 1,21	00:10:25.063 ± 00:01:31.744	00:01:37.382 ± 00:00:00.161
Unifei 2018	FD	512	79	73,92 ± 1,24	83,61 ± 2,27	85,61 ± 2,75	70,50 ± 2,39	77,82 ± 1,08	00:15:28.111 ± 00:04:11.211	00:02:44.329 ± 00:00:00.436
Unifei 2018	FD	1024	77	71,65 ± 3,30	79,98 ± 3,07	81,93 ± 4,94	67,45 ± 6,87	74,30 ± 1,42	00:14:57.735 ± 00:02:47.279	00:03:19.387 ± 00:00:00.688
Unifei 2018	MI	8	3	87,94 ± 1,04	69,21 ± 1,10	60,32 ± 2,15	91,49 ± 0,83	75,11 ± 1,15	00:00:53.613 ± 00:00:06.560	00:00:03.416 ± 00:00:00.179
Unifei 2018	MI	16	3	78,76 ± 1,49	76,47 ± 0,94	73,83 ± 1,85	80,89 ± 2,04	77,40 ± 0,60	00:00:55.755 ± 00:00:02.065	00:00:03.615 ± 00:00:00.030
Unifei 2018	MI	32	7	83,12 ± 1,22	76,73 ± 1,71	72,73 ± 3,16	85,73 ± 1,69	79,24 ± 1,04	00:01:46.418 ± 00:00:03.602	00:00:07.948 ± 00:00:00.026
Unifei 2018	MI	64	10	63,88 ± 0,47	81,41 ± 0,77	88,12 ± 0,83	51,00 ± 1,30	68,32 ± 0,52	00:04:00.895 ± 00:00:23.922	00:00:12.792 ± 00:00:00.034
Unifei 2018	MI	128	30	67,20 ± 0,98	78,84 ± 5,56	80,85 ± 11,96	61,22 ± 6,43	70,11 ± 3,30	00:05:19.331 ± 00:01:18.531	00:00:39.516 ± 00:00:00.324
Unifei 2018	MI	256	49	67,94 ± 3,14	78,14 ± 3,94	81,01 ± 8,11	61,51 ± 7,28	70,55 ± 1,01	00:13:20.290 ± 00:02:06.808	00:01:20.735 ± 00:00:00.221
Unifei 2018	MI	512	118	69,60 ± 1,75	76,13 ± 5,15	76,97 ± 8,53	66,97 ± 4,87	71,64 ± 2,40	00:22:03.625 ± 00:04:02.228	00:03:54.379 ± 00:00:00.465
Unifei 2018	MI	1024	155	74,35 ± 0,95	82,43 ± 5,36	82,79 ± 9,23	72,00 ± 3,80	77,07 ± 3,10	00:34:49.803 ± 00:05:55.031	00:06:14.633 ± 00:00:01.231

Tabela 22: Resultados do classificador REPT.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.004 ± 00:00:00.010	00:00:00.007 ± 00:00:00.008
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	77,42 ± 0,04	00:00:00.000 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	78,40 ± 0,03	00:00:00.000 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,19 ± 0,17	99,93 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.001 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,40 ± 0,09	99,91 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.000 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,73 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:00.003 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	512	14	68,66 ± 0,16	90,14 ± 0,36	97,08 ± 0,12	37,56 ± 0,48	69,05 ± 0,24	00:00:00.032 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	CHI2	1024	242	82,55 ± 0,22	91,70 ± 0,24	95,76 ± 0,15	69,84 ± 0,40	84,90 ± 0,19	00:00:03.029 ± 00:00:00.107	00:00:00.017 ± 00:00:00.000
Ling Spam	FD	8	4	98,27 ± 0,21	84,77 ± 0,15	80,64 ± 0,26	98,70 ± 0,16	89,95 ± 0,06	00:00:00.009 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	FD	16	4	98,33 ± 0,24	84,79 ± 0,14	80,67 ± 0,27	98,74 ± 0,18	89,99 ± 0,06	00:00:00.009 ± 00:00:00.000	00:00:00.002 ± 00:00:00.000
Ling Spam	FD	32	10	96,67 ± 2,68	93,20 ± 1,82	92,12 ± 2,30	96,87 ± 2,64	94,59 ± 0,34	00:00:00.025 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	64	19	98,99 ± 0,18	93,81 ± 0,20	92,86 ± 0,21	99,13 ± 0,16	96,13 ± 0,14	00:00:00.049 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	128	38	99,31 ± 0,15	95,47 ± 0,15	94,86 ± 0,19	99,39 ± 0,13	97,22 ± 0,08	00:00:00.108 ± 00:00:00.002	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	256	58	98,85 ± 0,99	96,22 ± 0,73	95,74 ± 0,85	98,95 ± 0,92	97,41 ± 0,19	00:00:00.154 ± 00:00:00.002	00:00:00.003 ± 00:00:00.000
Ling Spam	FD	512	103	97,58 ± 0,32	98,31 ± 0,15	98,16 ± 0,16	97,77 ± 0,29	97,96 ± 0,18	00:00:00.266 ± 00:00:00.008	00:00:00.004 ± 00:00:00.000
Ling Spam	FD	1024	121	98,44 ± 0,28	98,10 ± 0,14	97,92 ± 0,17	98,58 ± 0,25	98,26 ± 0,14	00:00:00.308 ± 00:00:00.010	00:00:00.004 ± 00:00:00.000
Ling Spam	MI	8	6	99,04 ± 0,33	85,05 ± 0,20	80,96 ± 0,27	99,28 ± 0,25	90,41 ± 0,10	00:00:00.014 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	16	12	99,22 ± 0,13	91,10 ± 0,23	89,41 ± 0,29	99,36 ± 0,11	94,58 ± 0,15	00:00:00.029 ± 00:00:00.000	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	32	19	93,38 ± 2,26	96,75 ± 1,82	96,43 ± 2,23	93,57 ± 2,23	94,93 ± 0,14	00:00:00.049 ± 00:00:00.001	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	64	34	96,30 ± 0,29	98,14 ± 0,21	98,00 ± 0,24	96,55 ± 0,26	97,24 ± 0,13	00:00:00.086 ± 00:00:00.002	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	128	52	96,23 ± 0,77	97,88 ± 0,69	97,70 ± 0,82	96,47 ± 0,73	97,06 ± 0,19	00:00:00.141 ± 00:00:00.003	00:00:00.003 ± 00:00:00.000
Ling Spam	MI	256	68	96,70 ± 0,57	98,04 ± 0,67	97,86 ± 0,80	96,92 ± 0,55	97,37 ± 0,15	00:00:00.187 ± 00:00:00.006	00:00:00.004 ± 00:00:00.000
Ling Spam	MI	512	139	97,77 ± 0,29	98,40 ± 0,17	98,26 ± 0,18	97,95 ± 0,27	98,10 ± 0,15	00:00:00.347 ± 00:00:00.011	00:00:00.004 ± 00:00:00.000
Ling Spam	MI	1024	157	97,64 ± 0,17	98,33 ± 0,15	98,19 ± 0,16	97,83 ± 0,14	98,00 ± 0,08	00:00:00.416 ± 00:00:00.017	00:00:00.005 ± 00:00:00.000
Spam Assassin	CHI2	8	4	67,55 ± 1,65	78,39 ± 2,34	79,91 ± 3,27	64,91 ± 4,12	71,85 ± 0,94	00:00:00.004 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	16	7	69,92 ± 0,93	78,34 ± 0,99	78,76 ± 1,59	69,27 ± 2,03	73,75 ± 0,49	00:00:00.007 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	32	14	71,83 ± 1,01	78,13 ± 0,71	77,57 ± 1,26	72,41 ± 1,71	74,86 ± 0,58	00:00:00.019 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	64	20	70,50 ± 1,29	81,31 ± 1,75	82,07 ± 2,44	69,14 ± 2,66	75,17 ± 0,60	00:00:00.031 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	128	25	74,20 ± 0,83	81,66 ± 1,12	81,53 ± 1,79	74,22 ± 1,73	77,69 ± 0,33	00:00:00.042 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	256	29	78,72 ± 0,80	89,42 ± 0,84	89,87 ± 0,99	77,81 ± 1,41	83,54 ± 0,65	00:00:00.044 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	512	7	62,43 ± 0,35	92,96 ± 0,58	96,06 ± 0,35	47,31 ± 0,36	68,89 ± 0,25	00:00:00.006 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	CHI2	1024	20	68,16 ± 0,33	88,85 ± 1,42	91,74 ± 1,37	60,31 ± 1,00	74,90 ± 0,28	00:00:00.017 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	8	5	71,22 ± 0,68	81,83 ± 0,82	83,26 ± 1,19	69,07 ± 1,27	75,79 ± 0,53	00:00:00.008 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	16	6	83,95 ± 0,90	89,24 ± 0,64	88,97 ± 0,86	84,26 ± 1,23	86,53 ± 0,51	00:00:00.007 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	32	9	86,84 ± 0,98	91,76 ± 0,95	91,47 ± 1,19	87,15 ± 1,28	89,23 ± 0,37	00:00:00.010 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	64	16	90,27 ± 0,79	92,55 ± 0,98	92,07 ± 1,09	90,82 ± 0,93	91,42 ± 0,50	00:00:00.018 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	128	26	92,81 ± 0,56	92,75 ± 0,43	92,09 ± 0,48	93,41 ± 0,54	92,77 ± 0,34	00:00:00.027 ± 00:00:00.001	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	256	53	93,52 ± 0,64	93,79 ± 0,62	93,25 ± 0,70	94,04 ± 0,60	93,66 ± 0,33	00:00:00.056 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	512	67	92,83 ± 1,51	91,85 ± 0,91	90,96 ± 1,22	93,44 ± 1,66	92,25 ± 0,58	00:00:00.072 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
Spam Assassin	FD	1024	58	88,78 ± 1,02	91,48 ± 0,90	90,97 ± 1,03	89,36 ± 1,09	90,13 ± 0,31	00:00:00.082 ± 00:00:00.003	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	8	3	80,20 ± 0,41	86,67 ± 0,48	86,49 ± 0,64	80,41 ± 0,76	83,32 ± 0,24	00:00:00.003 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	16	6	86,01 ± 1,25	91,63 ± 0,77	91,37 ± 0,96	86,31 ± 1,56	88,74 ± 0,50	00:00:00.006 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	32	18	91,46 ± 0,95	92,87 ± 0,56	92,35 ± 0,63	92,02 ± 1,00	92,17 ± 0,38	00:00:00.018 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	64	33	90,19 ± 0,57	93,91 ± 0,65	93,64 ± 0,68	90,59 ± 0,64	92,05 ± 0,30	00:00:00.035 ± 00:00:00.000	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	128	58	93,93 ± 0,83	94,16 ± 0,51	93,64 ± 0,62	94,40 ± 0,85	94,04 ± 0,35	00:00:00.064 ± 00:00:00.003	00:00:00.001 ± 00:00:00.000
Spam Assassin	MI	256	85	92,97 ± 0,90	93,92 ± 0,63	93,44 ± 0,70	93,47 ± 0,89	93,45 ± 0,25	00:00:00.091 ± 00:00:00.004	00:00:00.002 ± 00:00:00.000
Spam Assassin	MI	512	96	92,75 ± 0,54	93,67 ± 0,54	93,18 ± 0,58	93,27 ± 0,55	93,22 ± 0,26	00:00:00.101 ± 00:00:00.003	00:00:00.002 ± 00:00:00.000
Spam Assassin	MI	1024	84	92,07 ± 0,92	93,41 ± 0,45	92,91 ± 0,58	92,58 ± 1,00	92,74 ± 0,38	00:00:00.088 ± 00:00:00.002	00:00:00.001 ± 00:00:00.000
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	46,53 ± 0,01	00:00:00.000 ± 00:00:00.000	00:00:00.051 ± 00:00:00.004
TREC	CHI2	16	11	61,42 ± 0,00	90,69 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	47,31 ± 0,01	00:00:00.006 ± 00:00:00.004	00:00:00.062 ± 00:00:00.002
TREC	CHI2	32	10	61,39 ± 0,00	90,23 ± 0,73	99,94 ± 0,00	0,89 ± 0,02	47,23 ± 0,01	00:00:00.004 ± 00:00:00.000	00:00:00.062 ± 00:00:00.006
TREC	CHI2	64	12	64,70 ± 7,53	88,17 ± 12,37	90,01 ± 22,49	10,91 ± 22,38	44,80 ± 5,62	00:00:00.007 ± 00:00:00.000	00:00:00.067 ± 00:00:00.003
TREC	CHI2	128	45	61,70 ± 0,01	87,86 ± 0,21	99,84 ± 0,00	1,83 ± 0,02	48,14 ± 0,02	00:00:00.120 ± 00:00:00.011	00:00:00.100 ± 00:00:00.017
TREC	CHI2	256	68	84,51 ± 5,88	44,80 ± 12,24	11,69 ± 22,17	89,87 ± 21,91	26,71 ± 5,47	00:00:00.462 ± 00:00:00.011	00:00:00.184 ± 00:00:00.012
TREC	CHI2	512	133	61,58 ± 0,01	94,28 ± 0,21	99,79 ± 0,01	5,28 ± 0,06	49,92 ± 0,04	00:00:19.939 ± 00:00:00.356	00:00:00.362 ± 00:00:00.009
TREC	CHI2	1024	47	61,89 ± 0,04	77,00 ± 0,76	98,42 ± 0,09	7,99 ± 0,13	51,57 ± 0,08	00:00:07.657 ± 00:00:00.362	00:00:00.108 ± 00:00:00.007
TREC	FD	8	6	78,80 ± 0,32	83,45 ± 0,27	84,09 ± 0,38	78,00 ± 0,50	80,99 ± 0,09	00:00:00.633 ± 00:00:00.014	00:00:00.062 ± 00:00:00.002
TREC	FD	16	7	78,21 ± 0,18	84,57 ± 0,20	85,49 ± 0,26	76,94 ± 0,29	81,12 ± 0,09	00:00:00.772 ± 00:00:00.011	00:00:00.064 ± 00:00:00.001
TREC	FD	32	14	85,15 ± 0,14	89,48 ± 0,12	89,69 ± 0,14	84,86 ± 0,17	87,23 ± 0,05	00:00:01.530 ± 00:00:00.035	00:00:00.067 ± 00:00:00.001
TREC	FD	64	7	68,57 ± 0,10	91,00 ± 0,15	94,04 ± 0,12	58,29 ± 0,17	75,12 ± 0,06	00:00:00.935 ± 00:00:00.045	00:00:00.058 ± 00:00:00.006
TREC	FD	128	22	82,01 ± 0,19	91,22 ± 0,26	91,98 ± 0,27	80,50 ± 0,25	86,11 ± 0,10	00:00:03.221 ± 00:00:00.063	00:00:00.071 ± 00:00:00.003
TREC	FD	256	24	76,39 ± 0,13	93,08 ± 0,25	94,46 ± 0,22	71,86 ± 0,19	82,76 ± 0,07	00:00:04.986 ± 00:00:00.161	00:00:00.080 ± 00:00:00.005
TREC	FD	512	51	81,93 ± 0,22	92,33 ± 0,23	93,08 ± 0,24	80,22 ± 0,32	86,48 ± 0,12	00:00:16.006 ± 00:00:00.771	00:00:00.094 ± 00:00:00.005
TREC	FD	1024	49	76,48 ± 0,10	92,69 ± 0,16	94,07 ± 0,14	72,23 ± 0,14	82,75 ± 0,06	00:00:21.164 ± 00:00:00.924	00:00:00.100 ± 00:00:00.003
TREC	MI	8	5	81,87 ± 0,17	66,57 ± 0,07	57,73 ± 0,12	86,81 ± 0,15	71,48 ± 0,05	00:00:00.603 ± 00:00:00.068	00:00:00.057 ± 00:00:00.011
TREC	MI	16	6	82,46 ± 0,29	68,76 ± 0,09	60,15 ± 0,18	87,27 ± 0,28	73,25 ± 0,07	00:00:00.912 ± 00:00:00.053	00:00:00.054 ± 00:00:00.003
TREC	MI	32	9	65,76 ± 0,07	92,62 ± 0,16	95,72 ± 0,10	51,86 ± 0,11	72,13 ± 0,07	00:00:02.024 ± 00:00:00.138	00:00:00.061 ± 00:00:00.004
TREC	MI	64	16	73,52 ± 0,11	93,18 ± 0,15	94,95 ± 0,12	66,88 ± 0,12	80,33 ± 0,08	00:00:03.564 ± 00:00:00.143	00:00:00.063 ± 00:00:00.003
TREC	MI	128	26	79,38 ± 0,18	92,51 ± 0,26	93,61 ± 0,27	76,44 ± 0,33	84,79 ± 0,06	00:00:04.898 ± 00:00:00.202	00:00:00.074 ± 00:00:00.002
TREC	MI	256	37	76,00 ± 0,13	93,40 ± 0,14	94,81 ± 0,12	71,02 ± 0,18	82,50 ± 0,06	00:00:13.034 ± 00:00:00.813	00:00:00.086 ± 00:00:00.005
TREC	MI	512	66	79,81 ± 0,05	92,48 ± 0,10	93,53 ± 0,09	77,07 ± 0,09	85,09 ± 0,02	00:00:28.570 ± 00:00:00.690	00:00:00.104 ± 00:00:00.002
TREC	MI	1024	204	85,03 ± 0,17	93,88 ± 0,19	94,30 ± 0,20	84,04 ± 0,23	89,05 ± 0,06	00:01:30.832 ± 00:00:01.972	00:00:00.136 ± 00:00:00.003
Unifei 2017	CHI2	8	8	40,96 ± 0,00	96,48 ± 0,71	100,00 ± 0,00	0,09 ± 0,00	23,89 ± 0,00	00:00:00.012 ± 00:00:00.019	00:00:00.274 ± 00:00:00.029
Unifei 2017	CHI2	16	10	40,96 ± 0,00	88,56 ± 0,69	99,98 ± 0,00	0,13 ± 0,00	23,94 ± 0,00	00:00:00.017 ± 00:00:00.000	00:00:00.300 ± 00:00:00.009
Unifei 2017	CHI2	32	20	40,95 ± 0,00	84,96 ± 0,75	99,96 ± 0,00	0,17 ± 0,00	23,98 ± 0,00	00:00:00.061 ± 00:00:00.001	00:00:00.375 ± 00:00:00.011
Unifei 2017	CHI2	64	48	40,92 ± 0,00	84,13 ± 0,72	99,94 ± 0,00	0,21 ± 0,00	23,98 ± 0,00	00:00:00.310 ± 00:00:00.010	00:00:00.366 ± 00:00:00.023
Unifei 2017	CHI2	128	86	40,85 ± 0,00	85,56 ± 0,37	99,93 ± 0,00	0,28 ± 0,00	24,00 ± 0,00	00:00:01.169 ± 00:00:00.024	00:00:00.558 ± 00:00:00.160
Unifei 2017	CHI2	256	34	40,55 ± 0,00	68,24 ± 0,35	99,46 ± 0,01	0,78 ± 0,01	24,25 ± 0,01	00:00:00.429 ± 00:00:00.024	00:00:00.391 ± 00:00:00.038
Unifei 2017	CHI2	512	45	41,35 ± 0,01	64,79 ± 0,11	96,87 ± 0,01	4,02 ± 0,01	28,29 ± 0,01	00:00:05.827 ± 00:00:00.244	00:00:00.433 ± 00:00:00.076
Unifei 2017	CHI2	1024	17	44,98 ± 0,02	98,98 ± 0,04	99,90 ± 0,00	7,25 ± 0,02	34,45 ± 0,03	00:00:09.988 ± 00:00:00.753	00:00:00.276 ± 00:00:00.021
Unifei 2017	FD	8	5	89,32 ± 0,08	88,14 ± 0,09	87,92 ± 0,12	89,52 ± 0,10	88,72 ± 0,03	00:00:04.475 ± 00:00:00.081	00:00:00.303 ± 00:00:00.007
Unifei 2017	FD	16	3	68,40 ± 0,04	88,64 ± 0,12	92,64 ± 0,10	57,31 ± 0,12	74,15 ± 0,03	00:00:03.457 ± 00:00:00.203	00:00:00.238 ± 00:00:00.011
Unifei 2017	FD	32	6	79,05 ± 0,06	89,56 ± 0,12	91,13 ± 0,12	75,91 ± 0,11	83,41 ± 0,03	00:00:05.242 ± 00:00:00.154	00:00:00.294 ± 00:00:00.015
Unifei 2017	FD	64	22	89,85 ± 0,05	95,43 ± 0,06	95,71 ± 0,06	89,22 ± 0,05	92,45 ± 0,03	00:00:20.373 ± 00:00:00.446	00:00:00.311 ± 00:00:00.012
Unifei 2017	FD	128	65	90,72 ± 0,07	96,16 ± 0,06	96,39 ± 0,06	90,17 ± 0,08	93,27 ± 0,02	00:01:18.676 ± 00:00:01.901	00:00:00.356 ± 00:00:00.002
Unifei 2017	FD	256	51	87,99 ± 0,08	95,72 ± 0,12	96,10 ± 0,12	86,93 ± 0,11	91,49 ± 0,03	00:01:24.643 ± 00:00:01.564	00:00:00.374 ± 00:00:00.009
Unifei 2017	FD	512	101	89,59 ± 0,04	95,93 ± 0,04	96,22 ± 0,04	88,86 ± 0,04	92,52 ± 0,02	00:02:42.719 ± 00:00:07.158	00:00:00.425 ± 00:00:00.017
Unifei 2017	FD	1024	105	84,88 ± 0,06	95,91 ± 0,09	96,45 ± 0,08	82,86 ± 0,08	89,60 ± 0,03	00:05:26.909 ± 00:00:06.854	00:00:00.540 ± 00:00:00.017
Unifei 2017	MI	8	4	58,72 ± 1,93	85,06 ± 4,81	91,46 ± 5,84	38,39 ± 7,88	61,14 ± 2,29	00:00:02.728 ± 00:00:00.232	00:00:00.222 ± 00:00:00.002
Unifei 2017	MI	16	7	60,25 ± 0,03	93,41 ± 0,13	97,15 ± 0,06	38,63 ± 0,04	64,30 ± 0,03	00:00:08.128 ± 00:00:00.712	00:00:00.243 ± 00:00:00.012
Unifei 2017	MI	32	5	60,40 ± 0,03	94,08 ± 0,09	97,44 ± 0,04	38,87 ± 0,05	64,58 ± 0,03	00:00:03.832 ± 00:00:00.275	00:00:00.226 ± 00:00:00.028
Unifei 2017	MI	64	9	74,57 ± 0,06	89,55 ± 0,19	91,98 ± 0,18	68,65 ± 0,16	80,04 ± 0,02	00:00:08.951 ± 00:00:00.260	00:00:00.283 ± 00:00:00.017
Unifei 2017	MI	128	18	70,01 ± 0,04	94,86 ± 0,06	96,82 ± 0,04	58,57 ± 0,06	76,84 ± 0,03	00:00:35.697 ± 00:00:01.137	00:00:00.293 ± 00:00:00.005
Unifei 2017	MI	256	31	81,30 ± 0,07	94,10 ± 0,09	95,08 ± 0,08	78,19 ± 0,10	86,53 ± 0,04	00:00:58.875 ± 00:00:01.929	00:00:00.338 ± 00:00:00.018
Unifei 2017	MI	512	125	87,81 ± 0,03	96,15 ± 0,04	96,52 ± 0,04	86,64 ± 0,05	91,55 ± 0,02	00:06:06.142 ± 00:00:06.410	00:00:00.536 ± 00:00:00.010
Unifei 2017	MI	1024	208	89,44 ± 0,04	95,87 ± 0,07	96,16 ± 0,08	88,68 ± 0,05	92,40 ± 0,03	00:07:29.614 ± 00:00:10.041	00:00:00.581 ± 00:00:00.010
Unifei 2018	CHI2	8	4	99,80 ± 0,03	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	44,61 ± 0,00	00:00:00.012 ± 00:00:00.007	00:00:00.203 ± 00:00:00.002
Unifei 2018	CHI2	16	16	99,79 ± 0,03	59,17 ± 0,00	2,04 ± 0,01	100,00 ± 0,00	45,28 ± 0,00	00:00:00.056 ± 00:00:00.001	00:00:00.318 ± 00:00:00.022
Unifei 2018	CHI2	32	28	99,67 ± 0,07	59,46 ± 0,00	2,70 ± 0,01	99,99 ± 0,00	46,02 ± 0,00	00:00:00.232 ± 00:00:00.018	00:00:00.353 ± 00:00:00.035
Unifei 2018	CHI2	64	42	41,65 ± 0,00	99,04 ± 0,07	99,96 ± 0,00	2,57 ± 0,01	27,08 ± 0,01	00:00:00.624 ± 00:00:00.014	00:00:00.435 ± 00:00:00.025
Unifei 2018	CHI2	128	69	98,86 ± 0,02	59,96 ± 0,00	8,71 ± 0,02	99,93 ± 0,00	50,06 ± 0,01	00:00:04.898 ± 00:00:00.144	00:00:00.711 ± 00:00:00.015
Unifei 2018	CHI2	256	116	98,27 ± 0,04	60,53 ± 0,00	8,75 ± 0,02	99,89 ± 0,00	50,67 ± 0,01	00:00:13.714 ± 00:00:00.224	00:00:01.355 ± 00:00:00.093
Unifei 2018	CHI2	512	121	50,97 ± 0,01	95,80 ± 0,04	99,18 ± 0,01	16,38 ± 0,03	46,36 ± 0,02	00:03:13.493 ± 00:00:02.706	00:00:01.126 ± 00:00:00.049
Unifei 2018	CHI2	1024	152	54,15 ± 0,02	92,19 ± 0,07	97,76 ± 0,02	24,22 ± 0,04	53,34 ± 0,03	00:11:19.911 ± 00:00:14.234	00:00:01.172 ± 00:00:00.034
Unifei 2018	FD	8	5	96,23 ± 0,02	97,09 ± 0,04	97,07 ± 0,04	96,26 ± 0,02	96,66 ± 0,02	00:00:04.153 ± 00:00:00.014	00:00:00.293 ± 00:00:00.003
Unifei 2018	FD	16	3	72,59 ± 0,08	87,04 ± 0,21	89,87 ± 0,20	66,72 ± 0,16	77,90 ± 0,03	00:00:03.641 ± 00:00:00.126	00:00:00.259 ± 00:00:00.012
Unifei 2018	FD	32	8	86,73 ± 0,09	93,07 ± 0,12	93,42 ± 0,13	86,07 ± 0,12	89,69 ± 0,03	00:00:07.189 ± 00:00:00.191	00:00:00.298 ± 00:00:00.014
Unifei 2018	FD	64	11	75,23 ± 0,04	93,61 ± 0,13	95,13 ± 0,11	69,49 ± 0,09	81,86 ± 0,03	00:00:19.394 ± 00:00:01.415	00:00:00.300 ± 00:00:00.017
Unifei 2018	FD	128	24	85,38 ± 0,04	96,70 ± 0,03	97,06 ± 0,03	83,81 ± 0,04	90,32 ± 0,02	00:00:47.910 ± 00:00:01.278	00:00:00.328 ± 00:00:00.006
Unifei 2018	FD	256	60	89,55 ± 0,03	97,18 ± 0,07	97,35 ± 0,07	88,94 ± 0,06	93,08 ± 0,02	00:01:37.870 ± 00:00:01.994	00:00:00.385 ± 00:00:00.008
Unifei 2018	FD	512	79	89,54 ± 0,03	97,17 ± 0,04	97,34 ± 0,04	88,93 ± 0,04	93,07 ± 0,03	00:02:29.717 ± 00:00:03.890	00:00:00.420 ± 00:00:00.011
Unifei 2018	FD	1024	77	87,89 ± 0,05	95,95 ± 0,09	96,24 ± 0,08	87,06 ± 0,06	91,58 ± 0,02	00:02:52.172 ± 00:00:04.092	00:00:00.453 ± 00:00:00.015
Unifei 2018	MI	8	3	96,66 ± 0,03	72,13 ± 0,03	63,23 ± 0,05	97,76 ± 0,02	79,69 ± 0,03	00:00:02.095 ± 00:00:00.080	00:00:00.264 ± 00:00:00.003
Unifei 2018	MI	16	3	94,06 ± 0,06	88,76 ± 0,04	87,43 ± 0,05	94,73 ± 0,05	91,15 ± 0,02	00:00:02.513 ± 00:00:00.026	00:00:00.244 ± 00:00:00.011
Unifei 2018	MI	32	7	96,04 ± 0,45	90,57 ± 0,35	89,53 ± 0,47	96,45 ± 0,45	93,05 ± 0,02	00:00:05.551 ± 00:00:00.110	00:00:00.269 ± 00:00:00.012
Unifei 2018	MI	64	10	68,77 ± 0,04	95,37 ± 0,10	97,20 ± 0,07	56,62 ± 0,09	75,76 ± 0,04	00:00:17.499 ± 00:00:01.435	00:00:00.326 ± 00:00:00.117
Unifei 2018	MI	128	30	84,73 ± 0,08	93,74 ± 0,11	94,27 ± 0,12	83,48 ± 0,13	88,77 ± 0,04	00:00:43.881 ± 00:00:01.284	00:00:00.323 ± 00:00:00.008
Unifei 2018	MI	256	49	85,11 ± 0,05	96,26 ± 0,05	96,68 ± 0,05	83,50 ± 0,05	89,97 ± 0,02	00:01:25.362 ± 00:00:01.051	00:00:00.357 ± 00:00:00.008
Unifei 2018	MI	512	118	89,14 ± 0,08	96,98 ± 0,11	97,17 ± 0,11	88,47 ± 0,10	92,75 ± 0,02	00:03:15.757 ± 00:00:02.294	00:00:00.437 ± 00:00:00.006
Unifei 2018	MI	1024	155	88,76 ± 0,04	97,45 ± 0,06	97,64 ± 0,05	87,94 ± 0,04	92,72 ± 0,03	00:06:37.765 ± 00:00:05.401	00:00:00.546 ± 00:00:00.020

Tabela 23: Resultados do classificador AB-M1.

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	76,77 ± 0,03	00:00:00.006 ± 00:00:00.002	00:00:00.006 ± 00:00:00.001
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	77,42 ± 0,04	00:00:00.005 ± 00:00:00.001	00:00:00.005 ± 00:00:00.000
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	78,40 ± 0,03	00:00:00.005 ± 00:00:00.001	00:00:00.006 ± 00:00:00.000
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,16 ± 0,17	99,93 ± 0,01	28,14 ± 0,27	79,60 ± 0,03	00:00:00.007 ± 00:00:00.001	00:00:00.006 ± 00:00:00.000
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,40 ± 0,10	99,91 ± 0,01	40,44 ± 0,20	80,76 ± 0,02	00:00:00.014 ± 00:00:00.001	00:00:00.006 ± 00:00:00.000
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,72 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	81,71 ± 0,02	00:00:00.036 ± 00:00:00.004	00:00:00.006 ± 00:00:00.000
Ling Spam	CHI2	512	14	68,67 ± 0,15	90,18 ± 0,38	97,09 ± 0,13	37,60 ± 0,41	69,08 ± 0,20	00:00:00.333 ± 00:00:00.010	00:00:00.015 ± 00:00:00.001
Ling Spam	CHI2	1024	242	83,68 ± 0,22	92,99 ± 0,29	96,36 ± 0,16	71,99 ± 0,31	86,19 ± 0,18	00:00:20.305 ± 00:00:00.882	00:00:00.060 ± 00:00:00.005
Ling Spam	FD	8	4	98,73 ± 0,23	84,93 ± 0,13	80,82 ± 0,23	99,05 ± 0,18	90,22 ± 0,11	00:00:00.201 ± 00:00:00.009	00:00:00.014 ± 00:00:00.000
Ling Spam	FD	16	4	98,74 ± 0,25	84,99 ± 0,16	80,91 ± 0,28	99,05 ± 0,19	90,27 ± 0,11	00:00:00.210 ± 00:00:00.011	00:00:00.014 ± 00:00:00.001
Ling Spam	FD	32	10	95,72 ± 2,82	95,22 ± 2,08	94,53 ± 2,56	95,86 ± 2,78	95,22 ± 0,27	00:00:00.435 ± 00:00:00.018	00:00:00.014 ± 00:00:00.000
Ling Spam	FD	64	19	99,60 ± 0,16	94,80 ± 0,14	94,03 ± 0,16	99,66 ± 0,14	96,96 ± 0,12	00:00:00.764 ± 00:00:00.017	00:00:00.015 ± 00:00:00.000
Ling Spam	FD	128	38	99,79 ± 0,11	96,57 ± 0,08	96,13 ± 0,11	99,82 ± 0,09	98,05 ± 0,08	00:00:01.560 ± 00:00:00.020	00:00:00.017 ± 00:00:00.001
Ling Spam	FD	256	58	99,65 ± 0,59	97,58 ± 0,45	97,30 ± 0,51	99,67 ± 0,56	98,54 ± 0,07	00:00:02.375 ± 00:00:00.033	00:00:00.018 ± 00:00:00.001
Ling Spam	FD	512	103	99,43 ± 0,61	98,57 ± 0,38	98,42 ± 0,43	99,47 ± 0,57	98,97 ± 0,10	00:00:04.725 ± 00:00:00.188	00:00:00.018 ± 00:00:00.001
Ling Spam	FD	1024	121	99,08 ± 0,13	99,37 ± 0,12	99,32 ± 0,13	99,16 ± 0,11	99,23 ± 0,07	00:00:05.321 ± 00:00:00.249	00:00:00.018 ± 00:00:00.001
Ling Spam	MI	8	6	95,34 ± 5,76	87,91 ± 3,79	84,85 ± 5,04	95,06 ± 6,49	90,06 ± 1,01	00:00:00.256 ± 00:00:00.012	00:00:00.013 ± 00:00:00.000
Ling Spam	MI	16	12	99,50 ± 0,16	91,82 ± 0,24	90,31 ± 0,31	99,58 ± 0,14	95,13 ± 0,14	00:00:00.506 ± 00:00:00.014	00:00:00.015 ± 00:00:00.001
Ling Spam	MI	32	19	95,24 ± 2,75	96,33 ± 2,17	95,84 ± 2,60	95,38 ± 2,70	95,58 ± 0,19	00:00:00.785 ± 00:00:00.025	00:00:00.016 ± 00:00:00.001
Ling Spam	MI	64	34	96,92 ± 0,15	99,17 ± 0,11	99,11 ± 0,12	97,11 ± 0,12	98,07 ± 0,06	00:00:01.371 ± 00:00:00.038	00:00:00.016 ± 00:00:00.001
Ling Spam	MI	128	52	97,90 ± 1,21	98,18 ± 0,91	97,98 ± 1,04	98,03 ± 1,13	98,01 ± 0,15	00:00:02.209 ± 00:00:00.046	00:00:00.017 ± 00:00:00.000
Ling Spam	MI	256	68	97,75 ± 1,05	98,69 ± 0,86	98,55 ± 0,98	97,89 ± 0,99	98,20 ± 0,10	00:00:02.629 ± 00:00:00.081	00:00:00.016 ± 00:00:00.000
Ling Spam	MI	512	139	98,40 ± 0,13	99,34 ± 0,11	99,28 ± 0,12	98,52 ± 0,11	98,89 ± 0,06	00:00:05.618 ± 00:00:00.164	00:00:00.018 ± 00:00:00.001
Ling Spam	MI	1024	157	98,96 ± 0,54	98,95 ± 0,57	98,84 ± 0,65	99,04 ± 0,50	98,95 ± 0,07	00:00:06.903 ± 00:00:00.295	00:00:00.018 ± 00:00:00.000
Spam Assassin	CHI2	8	4	68,32 ± 0,89	77,54 ± 0,94	78,60 ± 1,38	66,89 ± 1,72	72,40 ± 0,42	00:00:00.057 ± 00:00:00.003	00:00:00.004 ± 00:00:00.000
Spam Assassin	CHI2	16	7	71,18 ± 0,68	78,31 ± 0,93	78,10 ± 1,29	71,38 ± 1,26	74,57 ± 0,31	00:00:00.103 ± 00:00:00.004	00:00:00.005 ± 00:00:00.000
Spam Assassin	CHI2	32	14	72,56 ± 0,51	78,81 ± 0,87	78,21 ± 1,25	73,22 ± 0,86	75,60 ± 0,51	00:00:00.245 ± 00:00:00.007	00:00:00.006 ± 00:00:00.000
Spam Assassin	CHI2	64	20	72,39 ± 1,34	80,50 ± 1,35	80,28 ± 2,10	72,50 ± 2,21	76,14 ± 0,42	00:00:00.414 ± 00:00:00.019	00:00:00.008 ± 00:00:00.000
Spam Assassin	CHI2	128	25	75,35 ± 1,50	82,47 ± 1,23	82,20 ± 1,92	75,47 ± 2,61	78,65 ± 0,57	00:00:00.552 ± 00:00:00.013	00:00:00.009 ± 00:00:00.000
Spam Assassin	CHI2	256	29	80,66 ± 0,56	91,24 ± 0,46	91,56 ± 0,53	79,98 ± 0,80	85,49 ± 0,45	00:00:00.596 ± 00:00:00.029	00:00:00.008 ± 00:00:00.000
Spam Assassin	CHI2	512	7	62,48 ± 0,37	93,13 ± 0,58	96,17 ± 0,35	47,37 ± 0,46	68,97 ± 0,29	00:00:00.086 ± 00:00:00.005	00:00:00.004 ± 00:00:00.000
Spam Assassin	CHI2	1024	20	68,95 ± 0,46	86,82 ± 2,47	89,55 ± 2,52	62,65 ± 1,40	75,20 ± 0,63	00:00:00.258 ± 00:00:00.016	00:00:00.005 ± 00:00:00.000
Spam Assassin	FD	8	5	74,81 ± 0,65	83,42 ± 0,80	83,94 ± 1,11	74,02 ± 1,04	78,75 ± 0,54	00:00:00.130 ± 00:00:00.003	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	16	6	88,63 ± 0,49	90,72 ± 0,52	90,10 ± 0,58	89,32 ± 0,54	89,70 ± 0,39	00:00:00.160 ± 00:00:00.004	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	32	9	90,92 ± 0,62	92,75 ± 0,45	92,26 ± 0,47	91,48 ± 0,68	91,86 ± 0,45	00:00:00.197 ± 00:00:00.005	00:00:00.007 ± 00:00:00.001
Spam Assassin	FD	64	16	93,67 ± 0,48	95,38 ± 0,28	95,06 ± 0,28	94,06 ± 0,51	94,54 ± 0,26	00:00:00.328 ± 00:00:00.007	00:00:00.006 ± 00:00:00.000
Spam Assassin	FD	128	26	95,90 ± 0,34	96,01 ± 0,32	95,67 ± 0,31	96,22 ± 0,32	95,96 ± 0,28	00:00:00.464 ± 00:00:00.012	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	256	53	96,88 ± 0,39	96,77 ± 0,36	96,49 ± 0,38	97,13 ± 0,37	96,82 ± 0,33	00:00:00.957 ± 00:00:00.017	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	512	67	96,48 ± 1,11	95,56 ± 0,70	95,11 ± 0,85	96,75 ± 1,11	95,96 ± 0,31	00:00:01.202 ± 00:00:00.026	00:00:00.007 ± 00:00:00.000
Spam Assassin	FD	1024	58	92,38 ± 0,55	94,16 ± 0,56	93,76 ± 0,59	92,85 ± 0,55	93,29 ± 0,40	00:00:01.326 ± 00:00:00.050	00:00:00.010 ± 00:00:00.000
Spam Assassin	MI	8	3	80,81 ± 0,67	87,25 ± 0,89	87,05 ± 1,10	81,04 ± 0,93	83,92 ± 0,23	00:00:00.078 ± 00:00:00.002	00:00:00.005 ± 00:00:00.000
Spam Assassin	MI	16	6	88,64 ± 0,82	91,96 ± 0,74	91,49 ± 0,83	89,23 ± 0,88	90,31 ± 0,30	00:00:00.132 ± 00:00:00.005	00:00:00.006 ± 00:00:00.000
Spam Assassin	MI	32	18	95,07 ± 0,58	94,82 ± 0,31	94,35 ± 0,32	95,48 ± 0,54	94,94 ± 0,25	00:00:00.326 ± 00:00:00.006	00:00:00.006 ± 00:00:00.000
Spam Assassin	MI	64	33	93,80 ± 0,53	96,08 ± 0,45	95,84 ± 0,48	94,15 ± 0,53	94,96 ± 0,32	00:00:00.570 ± 00:00:00.018	00:00:00.007 ± 00:00:00.000
Spam Assassin	MI	128	58	97,02 ± 0,38	96,93 ± 0,39	96,67 ± 0,41	97,25 ± 0,36	96,97 ± 0,29	00:00:00.955 ± 00:00:00.034	00:00:00.007 ± 00:00:00.000
Spam Assassin	MI	256	85	96,49 ± 0,39	97,28 ± 0,34	97,07 ± 0,35	96,74 ± 0,38	96,90 ± 0,24	00:00:01.503 ± 00:00:00.054	00:00:00.007 ± 00:00:00.000
Spam Assassin	MI	512	96	96,72 ± 0,31	96,50 ± 0,30	96,19 ± 0,33	96,98 ± 0,30	96,60 ± 0,20	00:00:01.666 ± 00:00:00.073	00:00:00.008 ± 00:00:00.000
Spam Assassin	MI	1024	84	96,88 ± 0,76	95,89 ± 0,58	95,48 ± 0,65	97,15 ± 0,71	96,35 ± 0,17	00:00:01.461 ± 00:00:00.030	00:00:00.008 ± 00:00:00.000
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	46,53 ± 0,01	00:00:00.006 ± 00:00:00.000	00:00:00.119 ± 00:00:00.005
TREC	CHI2	16	11	61,42 ± 0,00	90,69 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	47,31 ± 0,01	00:00:00.048 ± 00:00:00.000	00:00:00.190 ± 00:00:00.009
TREC	CHI2	32	10	61,39 ± 0,00	90,24 ± 0,73	99,94 ± 0,00	0,89 ± 0,02	47,23 ± 0,01	00:00:00.051 ± 00:00:00.004	00:00:00.182 ± 00:00:00.009
TREC	CHI2	64	12	61,37 ± 0,00	93,55 ± 0,39	99,96 ± 0,00	1,02 ± 0,01	47,28 ± 0,01	00:00:00.082 ± 00:00:00.002	00:00:00.220 ± 00:00:00.010
TREC	CHI2	128	45	61,70 ± 0,01	88,00 ± 0,20	99,84 ± 0,00	1,83 ± 0,02	48,14 ± 0,02	00:00:01.139 ± 00:00:00.053	00:00:00.570 ± 00:00:00.036
TREC	CHI2	256	68	77,18 ± 9,63	60,93 ± 19,90	41,08 ± 36,20	60,85 ± 35,77	33,99 ± 8,96	00:00:05.070 ± 00:00:00.188	00:00:00.965 ± 00:00:00.125
TREC	CHI2	512	133	89,39 ± 6,99	45,81 ± 12,24	13,11 ± 21,79	90,15 ± 21,32	29,15 ± 5,23	00:02:46.560 ± 00:00:03.144	00:00:01.825 ± 00:00:00.129
TREC	CHI2	1024	47	61,90 ± 0,03	77,29 ± 0,60	98,45 ± 0,07	8,00 ± 0,09	51,59 ± 0,06	00:01:07.350 ± 00:00:01.695	00:00:00.654 ± 00:00:00.039
TREC	FD	8	6	79,72 ± 0,31	84,79 ± 0,22	85,45 ± 0,30	78,86 ± 0,45	82,09 ± 0,11	00:00:10.040 ± 00:00:00.185	00:00:00.536 ± 00:00:00.041
TREC	FD	16	7	78,88 ± 0,10	85,85 ± 0,17	86,80 ± 0,20	77,51 ± 0,15	82,05 ± 0,06	00:00:11.929 ± 00:00:00.275	00:00:00.540 ± 00:00:00.044
TREC	FD	32	14	86,67 ± 0,21	91,28 ± 0,15	91,47 ± 0,18	86,39 ± 0,25	88,89 ± 0,08	00:00:22.563 ± 00:00:00.500	00:00:00.544 ± 00:00:00.024
TREC	FD	64	7	68,70 ± 0,11	91,46 ± 0,16	94,36 ± 0,13	58,40 ± 0,22	75,33 ± 0,08	00:00:15.497 ± 00:00:01.006	00:00:00.417 ± 00:00:00.035
TREC	FD	128	22	83,70 ± 0,17	92,21 ± 0,20	92,78 ± 0,22	82,55 ± 0,25	87,55 ± 0,06	00:00:44.904 ± 00:00:00.929	00:00:00.610 ± 00:00:00.018
TREC	FD	256	24	77,14 ± 0,10	93,80 ± 0,22	95,00 ± 0,19	72,86 ± 0,13	83,55 ± 0,07	00:01:25.494 ± 00:00:02.392	00:00:00.742 ± 00:00:00.057
TREC	FD	512	51	83,59 ± 0,13	93,56 ± 0,08	94,12 ± 0,09	82,20 ± 0,18	88,02 ± 0,07	00:03:51.299 ± 00:00:07.066	00:00:00.907 ± 00:00:00.029
TREC	FD	1024	49	77,29 ± 0,11	93,44 ± 0,11	94,64 ± 0,10	73,30 ± 0,15	83,59 ± 0,06	00:05:18.728 ± 00:00:10.739	00:00:01.126 ± 00:00:00.074
TREC	MI	8	5	82,22 ± 0,17	66,68 ± 0,07	57,81 ± 0,10	87,11 ± 0,16	71,65 ± 0,05	00:00:08.373 ± 00:00:00.377	00:00:00.282 ± 00:00:00.016
TREC	MI	16	6	82,96 ± 0,22	68,92 ± 0,06	60,27 ± 0,16	87,68 ± 0,22	73,51 ± 0,05	00:00:12.237 ± 00:00:00.521	00:00:00.334 ± 00:00:00.016
TREC	MI	32	9	65,75 ± 0,08	93,02 ± 0,17	95,98 ± 0,11	51,70 ± 0,18	72,15 ± 0,08	00:00:25.794 ± 00:00:01.505	00:00:00.409 ± 00:00:00.025
TREC	MI	64	16	73,57 ± 0,14	93,84 ± 0,16	95,47 ± 0,13	66,78 ± 0,17	80,53 ± 0,10	00:01:01.046 ± 00:00:02.530	00:00:00.530 ± 00:00:00.026
TREC	MI	128	26	80,09 ± 0,23	93,32 ± 0,35	94,29 ± 0,35	77,29 ± 0,38	85,56 ± 0,05	00:01:08.365 ± 00:00:01.224	00:00:00.622 ± 00:00:00.015
TREC	MI	256	37	76,43 ± 0,12	94,10 ± 0,20	95,37 ± 0,17	71,54 ± 0,19	83,04 ± 0,06	00:03:29.475 ± 00:00:06.597	00:00:00.970 ± 00:00:00.064
TREC	MI	512	66	80,51 ± 0,08	93,78 ± 0,10	94,68 ± 0,09	77,79 ± 0,13	86,02 ± 0,05	00:07:57.123 ± 00:00:09.733	00:00:01.529 ± 00:00:00.037
TREC	MI	1024	204	86,64 ± 0,14	95,37 ± 0,12	95,67 ± 0,12	85,81 ± 0,16	90,63 ± 0,07	00:19:21.268 ± 00:00:29.870	00:00:01.610 ± 00:00:00.104
Unifei 2017	CHI2	8	8	40,96 ± 0,00	96,44 ± 0,52	100,00 ± 0,00	0,09 ± 0,00	23,89 ± 0,00	00:00:00.020 ± 00:00:00.004	00:00:00.584 ± 00:00:00.026
Unifei 2017	CHI2	16	10	40,96 ± 0,00	88,57 ± 0,70	99,98 ± 0,00	0,13 ± 0,00	23,94 ± 0,00	00:00:00.184 ± 00:00:00.002	00:00:00.936 ± 00:00:00.043
Unifei 2017	CHI2	32	20	40,95 ± 0,00	84,98 ± 0,75	99,96 ± 0,00	0,17 ± 0,00	23,98 ± 0,00	00:00:00.533 ± 00:00:00.032	00:00:01.366 ± 00:00:00.063
Unifei 2017	CHI2	64	48	40,92 ± 0,00	84,11 ± 0,71	99,94 ± 0,00	0,21 ± 0,00	23,98 ± 0,00	00:00:02.182 ± 00:00:00.131	00:00:02.430 ± 00:00:00.211
Unifei 2017	CHI2	128	86	45,78 ± 11,15	82,96 ± 5,96	89,97 ± 22,52	10,25 ± 22,55	26,03 ± 4,60	00:00:08.003 ± 00:00:00.237	00:00:03.052 ± 00:00:00.503
Unifei 2017	CHI2	256	34	40,56 ± 0,00	68,29 ± 0,36	99,46 ± 0,01	0,79 ± 0,01	24,26 ± 0,01	00:00:04.381 ± 00:00:00.101	00:00:02.474 ± 00:00:00.246
Unifei 2017	CHI2	512	45	41,35 ± 0,01	64,83 ± 0,10	96,88 ± 0,01	4,02 ± 0,01	28,29 ± 0,01	00:00:43.125 ± 00:00:03.333	00:00:02.481 ± 00:00:00.244
Unifei 2017	CHI2	1024	17	44,98 ± 0,02	98,97 ± 0,04	99,90 ± 0,00	7,25 ± 0,02	34,45 ± 0,03	00:01:20.912 ± 00:00:04.791	00:00:00.856 ± 00:00:00.047
Unifei 2017	FD	8	5	90,44 ± 0,18	88,39 ± 0,18	88,05 ± 0,24	90,72 ± 0,21	89,38 ± 0,03	00:00:59.663 ± 00:00:00.960	00:00:02.519 ± 00:00:00.052
Unifei 2017	FD	16	3	68,59 ± 0,05	88,66 ± 0,18	92,60 ± 0,15	57,69 ± 0,17	74,34 ± 0,03	00:00:39.823 ± 00:00:01.454	00:00:01.410 ± 00:00:00.040
Unifei 2017	FD	32	6	79,73 ± 0,05	89,90 ± 0,12	91,35 ± 0,12	76,83 ± 0,10	84,00 ± 0,02	00:01:10.422 ± 00:00:01.092	00:00:02.233 ± 00:00:00.028
Unifei 2017	FD	64	22	90,91 ± 0,11	95,66 ± 0,16	95,89 ± 0,17	90,44 ± 0,14	93,16 ± 0,03	00:04:23.635 ± 00:00:06.434	00:00:02.588 ± 00:00:00.054
Unifei 2017	FD	128	65	91,56 ± 0,07	96,66 ± 0,11	96,84 ± 0,11	91,10 ± 0,09	93,96 ± 0,02	00:17:11.884 ± 00:00:38.035	00:00:03.276 ± 00:00:00.121
Unifei 2017	FD	256	51	88,83 ± 0,05	96,00 ± 0,11	96,33 ± 0,11	87,92 ± 0,08	92,11 ± 0,02	00:20:01.866 ± 00:00:32.277	00:00:04.252 ± 00:00:00.189
Unifei 2017	FD	512	101	90,57 ± 0,16	96,06 ± 0,26	96,30 ± 0,26	90,00 ± 0,22	93,14 ± 0,03	00:31:32.995 ± 00:00:35.074	00:00:03.702 ± 00:00:00.093
Unifei 2017	FD	1024	105	85,53 ± 0,05	96,34 ± 0,10	96,81 ± 0,09	83,67 ± 0,07	90,19 ± 0,03	00:57:03.542 ± 00:01:09.837	00:00:06.905 ± 00:00:00.472
Unifei 2017	MI	8	4	57,48 ± 0,04	88,46 ± 0,11	95,42 ± 0,06	33,22 ± 0,06	59,70 ± 0,04	00:00:24.061 ± 00:00:03.132	00:00:00.661 ± 00:00:00.099
Unifei 2017	MI	16	7	60,32 ± 0,03	93,42 ± 0,13	97,15 ± 0,06	38,83 ± 0,08	64,43 ± 0,04	00:01:01.996 ± 00:00:07.580	00:00:00.876 ± 00:00:00.103
Unifei 2017	MI	32	5	60,47 ± 0,03	94,36 ± 0,09	97,56 ± 0,04	38,97 ± 0,05	64,69 ± 0,03	00:00:37.254 ± 00:00:03.649	00:00:00.694 ± 00:00:00.066
Unifei 2017	MI	64	9	74,94 ± 0,06	89,31 ± 0,19	91,69 ± 0,18	69,36 ± 0,15	80,28 ± 0,03	00:01:19.022 ± 00:00:05.980	00:00:01.440 ± 00:00:00.156
Unifei 2017	MI	128	18	70,20 ± 0,04	94,94 ± 0,10	96,86 ± 0,07	58,93 ± 0,06	77,06 ± 0,03	00:05:07.324 ± 00:00:29.872	00:00:01.647 ± 00:00:00.218
Unifei 2017	MI	256	31	81,77 ± 0,07	94,29 ± 0,07	95,21 ± 0,06	78,84 ± 0,07	86,93 ± 0,04	00:10:49.084 ± 00:00:16.762	00:00:03.494 ± 00:00:00.251
Unifei 2017	MI	512	125	88,43 ± 0,04	96,55 ± 0,06	96,87 ± 0,05	87,37 ± 0,05	92,10 ± 0,02	01:09:28.911 ± 00:01:03.342	00:00:07.712 ± 00:00:00.440
Unifei 2017	MI	1024	208	90,12 ± 0,07	96,31 ± 0,09	96,56 ± 0,09	89,45 ± 0,09	92,99 ± 0,02	01:20:43.963 ± 00:01:55.699	00:00:05.679 ± 00:00:00.312
Unifei 2018	CHI2	8	4	99,81 ± 0,05	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	44,61 ± 0,00	00:00:00.052 ± 00:00:00.013	00:00:00.470 ± 00:00:00.037
Unifei 2018	CHI2	16	16	99,89 ± 0,05	59,17 ± 0,00	2,05 ± 0,01	100,00 ± 0,00	45,28 ± 0,00	00:00:00.500 ± 00:00:00.021	00:00:01.037 ± 00:00:00.067
Unifei 2018	CHI2	32	28	99,82 ± 0,04	59,46 ± 0,00	2,70 ± 0,01	100,00 ± 0,00	46,02 ± 0,00	00:00:01.760 ± 00:00:00.154	00:00:01.684 ± 00:00:00.200
Unifei 2018	CHI2	64	42	41,65 ± 0,00	99,09 ± 0,07	99,97 ± 0,00	2,57 ± 0,01	27,08 ± 0,01	00:00:06.815 ± 00:00:00.472	00:00:02.389 ± 00:00:00.253
Unifei 2018	CHI2	128	69	98,86 ± 0,02	59,96 ± 0,00	8,71 ± 0,02	99,93 ± 0,00	50,06 ± 0,01	00:00:29.269 ± 00:00:02.029	00:00:02.066 ± 00:00:00.317
Unifei 2018	CHI2	256	116	98,31 ± 0,05	60,53 ± 0,00	8,76 ± 0,02	99,89 ± 0,00	50,69 ± 0,01	00:02:02.954 ± 00:00:13.535	00:00:04.115 ± 00:00:00.582
Unifei 2018	CHI2	512	121	50,98 ± 0,01	95,81 ± 0,06	99,18 ± 0,01	16,39 ± 0,02	46,37 ± 0,02	00:17:50.340 ± 00:01:26.506	00:00:04.044 ± 00:00:00.320
Unifei 2018	CHI2	1024	152	54,16 ± 0,02	92,18 ± 0,06	97,75 ± 0,02	24,25 ± 0,05	53,36 ± 0,03	00:59:44.921 ± 00:04:41.307	00:00:07.062 ± 00:00:00.743
Unifei 2018	FD	8	5	98,04 ± 0,04	97,56 ± 0,03	97,51 ± 0,04	98,08 ± 0,03	97,80 ± 0,03	00:00:39.308 ± 00:00:01.229	00:00:02.491 ± 00:00:00.066
Unifei 2018	FD	16	3	72,74 ± 0,07	87,14 ± 0,20	89,92 ± 0,19	66,95 ± 0,16	78,05 ± 0,03	00:00:33.763 ± 00:00:02.489	00:00:01.357 ± 00:00:00.128
Unifei 2018	FD	32	8	87,47 ± 0,22	93,86 ± 0,27	94,17 ± 0,30	86,85 ± 0,31	90,45 ± 0,03	00:00:58.675 ± 00:00:00.876	00:00:02.326 ± 00:00:00.085
Unifei 2018	FD	64	11	75,57 ± 0,05	93,83 ± 0,11	95,27 ± 0,09	70,01 ± 0,10	82,21 ± 0,04	00:03:16.049 ± 00:00:19.765	00:00:02.084 ± 00:00:00.266
Unifei 2018	FD	128	24	86,07 ± 0,03	97,06 ± 0,04	97,37 ± 0,04	84,64 ± 0,04	90,89 ± 0,02	00:08:25.402 ± 00:00:13.562	00:00:03.653 ± 00:00:00.134
Unifei 2018	FD	256	60	90,45 ± 0,07	97,73 ± 0,10	97,86 ± 0,09	89,93 ± 0,08	93,84 ± 0,02	00:13:55.128 ± 00:00:18.012	00:00:03.970 ± 00:00:00.117
Unifei 2018	FD	512	79	90,47 ± 0,03	97,83 ± 0,03	97,95 ± 0,03	89,95 ± 0,04	93,89 ± 0,03	00:19:48.020 ± 00:00:18.148	00:00:04.125 ± 00:00:00.132
Unifei 2018	FD	1024	77	88,83 ± 0,07	96,58 ± 0,10	96,80 ± 0,10	88,13 ± 0,09	92,40 ± 0,02	00:16:21.321 ± 00:00:23.137	00:00:04.075 ± 00:00:00.073
Unifei 2018	MI	8	3	96,75 ± 0,02	72,14 ± 0,03	63,22 ± 0,05	97,82 ± 0,02	79,71 ± 0,03	00:00:30.073 ± 00:00:02.852	00:00:01.062 ± 00:00:00.121
Unifei 2018	MI	16	3	94,16 ± 0,07	88,84 ± 0,04	87,51 ± 0,05	94,82 ± 0,07	91,23 ± 0,05	00:00:27.871 ± 00:00:01.616	00:00:01.382 ± 00:00:00.118
Unifei 2018	MI	32	7	96,38 ± 0,44	90,82 ± 0,35	89,80 ± 0,46	96,76 ± 0,43	93,34 ± 0,03	00:01:03.509 ± 00:00:02.439	00:00:01.818 ± 00:00:00.037
Unifei 2018	MI	64	10	68,95 ± 0,03	95,58 ± 0,06	97,32 ± 0,04	56,92 ± 0,08	75,99 ± 0,03	00:02:22.183 ± 00:00:20.620	00:00:01.226 ± 00:00:00.222
Unifei 2018	MI	128	30	85,53 ± 0,04	94,01 ± 0,06	94,47 ± 0,06	84,45 ± 0,05	89,37 ± 0,03	00:06:29.195 ± 00:00:12.472	00:00:02.944 ± 00:00:00.083
Unifei 2018	MI	256	49	85,87 ± 0,06	96,61 ± 0,09	96,96 ± 0,08	84,44 ± 0,07	90,59 ± 0,02	00:14:59.339 ± 00:00:19.900	00:00:04.121 ± 00:00:00.210
Unifei 2018	MI	512	118	90,11 ± 0,07	97,51 ± 0,09	97,65 ± 0,09	89,55 ± 0,10	93,54 ± 0,02	00:31:53.224 ± 00:00:34.140	00:00:04.714 ± 00:00:00.261
Unifei 2018	MI	1024	155	89,64 ± 0,04	97,95 ± 0,05	98,09 ± 0,04	88,94 ± 0,04	93,45 ± 0,02	01:08:49.197 ± 00:01:50.739	00:00:06.498 ± 00:00:00.336

Tabela 24: Resultados do classificador SVM (linear).

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	24,99 ± 28,79	40,33 ± 26,35	29,98 ± 34,53	71,63 ± 32,68	30,21 ± 17,69	00:00:00.044 ± 00:00:00.002	00:00:00.049 ± 00:00:00.001
Ling Spam	CHI2	16	1	49,90 ± 30,72	65,17 ± 28,82	59,95 ± 36,91	45,90 ± 33,31	47,08 ± 19,43	00:00:00.045 ± 00:00:00.002	00:00:00.049 ± 00:00:00.001
Ling Spam	CHI2	32	1	41,43 ± 31,24	59,07 ± 29,44	49,96 ± 37,67	58,73 ± 31,12	43,94 ± 20,55	00:00:00.044 ± 00:00:00.001	00:00:00.050 ± 00:00:00.001
Ling Spam	CHI2	64	3	89,39 ± 6,53	68,67 ± 28,10	60,08 ± 36,80	56,91 ± 26,53	67,15 ± 7,32	00:00:00.044 ± 00:00:00.001	00:00:00.048 ± 00:00:00.001
Ling Spam	CHI2	128	5	88,82 ± 6,88	70,82 ± 26,44	60,21 ± 36,67	64,26 ± 22,01	69,94 ± 8,38	00:00:00.045 ± 00:00:00.001	00:00:00.048 ± 00:00:00.001
Ling Spam	CHI2	256	14	91,97 ± 7,41	60,62 ± 24,05	41,33 ± 36,07	80,62 ± 17,89	67,74 ± 8,89	00:00:00.050 ± 00:00:00.001	00:00:00.053 ± 00:00:00.002
Ling Spam	CHI2	512	14	83,13 ± 11,28	65,35 ± 17,68	49,40 ± 35,59	68,87 ± 23,41	65,73 ± 5,12	00:00:00.057 ± 00:00:00.003	00:00:00.055 ± 00:00:00.001
Ling Spam	CHI2	1024	242	91,58 ± 5,98	69,73 ± 17,07	60,57 ± 26,70	86,01 ± 10,41	76,74 ± 6,99	00:00:00.247 ± 00:00:00.004	00:00:00.159 ± 00:00:00.001
Ling Spam	FD	8	4	96,25 ± 0,27	78,14 ± 0,39	70,23 ± 0,67	97,49 ± 0,18	85,50 ± 0,21	00:00:00.055 ± 00:00:00.001	00:00:00.052 ± 00:00:00.001
Ling Spam	FD	16	4	96,26 ± 0,27	78,25 ± 0,38	70,43 ± 0,65	97,49 ± 0,18	85,58 ± 0,20	00:00:00.055 ± 00:00:00.002	00:00:00.051 ± 00:00:00.001
Ling Spam	FD	32	10	94,28 ± 2,91	90,65 ± 1,78	89,17 ± 2,44	94,80 ± 2,86	92,22 ± 0,40	00:00:00.065 ± 00:00:00.002	00:00:00.058 ± 00:00:00.001
Ling Spam	FD	64	19	97,38 ± 0,18	91,74 ± 0,21	90,39 ± 0,26	97,77 ± 0,16	94,32 ± 0,14	00:00:00.080 ± 00:00:00.002	00:00:00.067 ± 00:00:00.001
Ling Spam	FD	128	38	95,09 ± 2,52	95,99 ± 0,90	95,61 ± 1,09	95,34 ± 2,51	95,51 ± 0,77	00:00:00.106 ± 00:00:00.002	00:00:00.083 ± 00:00:00.001
Ling Spam	FD	256	58	98,85 ± 0,15	95,80 ± 0,20	95,27 ± 0,24	98,98 ± 0,14	97,22 ± 0,12	00:00:00.148 ± 00:00:00.002	00:00:00.097 ± 00:00:00.001
Ling Spam	FD	512	103	98,18 ± 0,75	98,01 ± 0,48	97,81 ± 0,53	98,33 ± 0,69	98,08 ± 0,13	00:00:00.214 ± 00:00:00.003	00:00:00.136 ± 00:00:00.003
Ling Spam	FD	1024	121	98,11 ± 0,20	98,22 ± 0,11	98,06 ± 0,12	98,27 ± 0,18	98,16 ± 0,08	00:00:00.250 ± 00:00:00.005	00:00:00.145 ± 00:00:00.002
Ling Spam	MI	8	6	81,40 ± 4,43	94,44 ± 2,77	94,45 ± 3,88	79,46 ± 4,98	87,43 ± 0,69	00:00:00.058 ± 00:00:00.002	00:00:00.054 ± 00:00:00.002
Ling Spam	MI	16	12	94,46 ± 4,13	92,64 ± 2,31	91,50 ± 3,00	94,59 ± 4,18	93,30 ± 0,76	00:00:00.066 ± 00:00:00.000	00:00:00.060 ± 00:00:00.001
Ling Spam	MI	32	19	95,97 ± 2,71	93,32 ± 2,00	92,31 ± 2,52	96,23 ± 2,63	94,46 ± 0,23	00:00:00.076 ± 00:00:00.001	00:00:00.067 ± 00:00:00.002
Ling Spam	MI	64	34	97,30 ± 1,09	95,36 ± 1,39	94,76 ± 1,64	97,56 ± 1,03	96,24 ± 0,28	00:00:00.100 ± 00:00:00.002	00:00:00.077 ± 00:00:00.001
Ling Spam	MI	128	52	97,72 ± 1,19	96,55 ± 0,94	96,15 ± 1,09	97,91 ± 1,12	97,08 ± 0,14	00:00:00.132 ± 00:00:00.004	00:00:00.092 ± 00:00:00.001
Ling Spam	MI	256	68	97,95 ± 1,15	96,93 ± 0,94	96,58 ± 1,08	98,12 ± 1,08	97,39 ± 0,11	00:00:00.154 ± 00:00:00.002	00:00:00.104 ± 00:00:00.001
Ling Spam	MI	512	139	98,96 ± 0,60	97,07 ± 0,67	96,73 ± 0,76	99,06 ± 0,54	97,96 ± 0,09	00:00:00.266 ± 00:00:00.003	00:00:00.159 ± 00:00:00.001
Ling Spam	MI	1024	157	97,33 ± 0,21	98,65 ± 0,13	98,55 ± 0,14	97,53 ± 0,18	98,01 ± 0,08	00:00:00.305 ± 00:00:00.007	00:00:00.171 ± 00:00:00.001
Spam Assassin	CHI2	8	4	62,19 ± 1,58	86,15 ± 2,48	90,86 ± 2,83	49,63 ± 4,29	72,14 ± 0,44	00:00:00.048 ± 00:00:00.002	00:00:00.046 ± 00:00:00.002
Spam Assassin	CHI2	16	7	63,69 ± 1,28	83,99 ± 2,88	88,12 ± 3,09	54,37 ± 3,86	72,51 ± 0,61	00:00:00.048 ± 00:00:00.001	00:00:00.048 ± 00:00:00.002
Spam Assassin	CHI2	32	14	64,29 ± 1,33	86,66 ± 2,78	90,33 ± 3,04	54,35 ± 4,02	73,87 ± 0,36	00:00:00.052 ± 00:00:00.001	00:00:00.051 ± 00:00:00.001
Spam Assassin	CHI2	64	20	67,17 ± 1,26	83,42 ± 1,18	86,08 ± 1,50	62,25 ± 2,56	74,72 ± 0,52	00:00:00.056 ± 00:00:00.001	00:00:00.052 ± 00:00:00.002
Spam Assassin	CHI2	128	25	68,29 ± 1,43	86,06 ± 2,01	88,55 ± 2,53	62,47 ± 3,72	76,33 ± 0,30	00:00:00.059 ± 00:00:00.002	00:00:00.055 ± 00:00:00.002
Spam Assassin	CHI2	256	29	75,68 ± 0,68	89,73 ± 0,95	90,76 ± 1,07	73,40 ± 1,01	82,39 ± 0,54	00:00:00.063 ± 00:00:00.002	00:00:00.054 ± 00:00:00.001
Spam Assassin	CHI2	512	7	64,80 ± 5,12	88,64 ± 8,95	86,55 ± 20,96	52,86 ± 11,73	72,98 ± 3,54	00:00:00.050 ± 00:00:00.002	00:00:00.048 ± 00:00:00.001
Spam Assassin	CHI2	1024	20	63,92 ± 0,57	85,32 ± 2,13	90,08 ± 1,66	52,91 ± 1,30	73,02 ± 0,71	00:00:00.057 ± 00:00:00.001	00:00:00.053 ± 00:00:00.001
Spam Assassin	FD	8	5	64,81 ± 0,58	71,06 ± 4,65	69,87 ± 7,53	65,19 ± 3,59	67,73 ± 2,17	00:00:00.049 ± 00:00:00.002	00:00:00.048 ± 00:00:00.000
Spam Assassin	FD	16	6	78,20 ± 0,90	78,24 ± 4,20	74,98 ± 6,19	80,63 ± 2,17	78,01 ± 2,00	00:00:00.049 ± 00:00:00.001	00:00:00.047 ± 00:00:00.001
Spam Assassin	FD	32	9	81,75 ± 0,70	74,95 ± 4,59	67,98 ± 7,53	86,08 ± 1,14	77,68 ± 2,89	00:00:00.052 ± 00:00:00.002	00:00:00.051 ± 00:00:00.002
Spam Assassin	FD	64	16	84,91 ± 1,03	83,19 ± 3,13	80,71 ± 4,23	86,72 ± 1,36	83,88 ± 1,53	00:00:00.058 ± 00:00:00.002	00:00:00.054 ± 00:00:00.002
Spam Assassin	FD	128	26	92,22 ± 0,60	90,65 ± 1,01	89,59 ± 1,19	93,01 ± 0,60	91,37 ± 0,31	00:00:00.063 ± 00:00:00.001	00:00:00.056 ± 00:00:00.002
Spam Assassin	FD	256	53	94,61 ± 0,66	93,88 ± 0,85	93,28 ± 0,95	95,09 ± 0,63	94,22 ± 0,37	00:00:00.082 ± 00:00:00.002	00:00:00.065 ± 00:00:00.001
Spam Assassin	FD	512	67	93,28 ± 1,22	94,36 ± 0,94	93,92 ± 1,07	93,72 ± 1,24	93,82 ± 0,35	00:00:00.091 ± 00:00:00.001	00:00:00.069 ± 00:00:00.001
Spam Assassin	FD	1024	58	87,37 ± 0,88	82,01 ± 2,73	78,49 ± 3,91	89,49 ± 1,11	84,34 ± 1,32	00:00:00.090 ± 00:00:00.002	00:00:00.066 ± 00:00:00.001
Spam Assassin	MI	8	3	33,10 ± 30,57	61,56 ± 8,87	27,85 ± 25,72	94,74 ± 4,86	52,03 ± 16,47	00:00:00.048 ± 00:00:00.002	00:00:00.047 ± 00:00:00.001
Spam Assassin	MI	16	6	97,06 ± 2,57	68,43 ± 5,98	48,44 ± 13,64	98,01 ± 1,79	77,57 ± 4,06	00:00:00.050 ± 00:00:00.002	00:00:00.048 ± 00:00:00.001
Spam Assassin	MI	32	18	87,46 ± 0,67	90,13 ± 0,71	89,54 ± 0,77	88,14 ± 0,71	88,82 ± 0,29	00:00:00.061 ± 00:00:00.001	00:00:00.054 ± 00:00:00.001
Spam Assassin	MI	64	33	90,52 ± 1,03	84,82 ± 3,23	81,83 ± 4,46	92,01 ± 1,28	87,29 ± 1,45	00:00:00.067 ± 00:00:00.001	00:00:00.061 ± 00:00:00.002
Spam Assassin	MI	128	58	93,23 ± 0,45	96,77 ± 0,52	96,62 ± 0,53	93,53 ± 0,42	95,04 ± 0,28	00:00:00.087 ± 00:00:00.001	00:00:00.071 ± 00:00:00.003
Spam Assassin	MI	256	85	95,01 ± 0,47	94,20 ± 0,88	93,63 ± 0,98	95,46 ± 0,45	94,58 ± 0,36	00:00:00.104 ± 00:00:00.004	00:00:00.077 ± 00:00:00.002
Spam Assassin	MI	512	96	94,22 ± 0,56	93,67 ± 0,67	93,07 ± 0,73	94,73 ± 0,54	93,92 ± 0,29	00:00:00.109 ± 00:00:00.001	00:00:00.079 ± 00:00:00.001
Spam Assassin	MI	1024	84	94,03 ± 0,83	93,56 ± 0,81	92,94 ± 0,92	94,54 ± 0,80	93,77 ± 0,24	00:00:00.101 ± 00:00:00.002	00:00:00.075 ± 00:00:00.002
TREC	CHI2	8	4	80,01 ± 14,26	68,24 ± 22,05	50,06 ± 37,66	50,15 ± 37,59	59,72 ± 1,30	00:00:00.044 ± 00:00:00.001	00:00:00.118 ± 00:00:00.001
TREC	CHI2	16	11	78,86 ± 13,15	64,78 ± 19,52	50,20 ± 37,50	50,47 ± 37,32	59,15 ± 1,15	00:00:00.047 ± 00:00:00.002	00:00:00.118 ± 00:00:00.001
TREC	CHI2	32	10	89,17 ± 10,47	49,09 ± 15,37	20,32 ± 30,02	80,15 ± 29,89	58,16 ± 0,88	00:00:00.048 ± 00:00:00.001	00:00:00.119 ± 00:00:00.003
TREC	CHI2	64	12	81,84 ± 12,61	60,77 ± 20,13	40,31 ± 36,72	60,38 ± 36,55	58,96 ± 1,33	00:00:00.048 ± 00:00:00.001	00:00:00.119 ± 00:00:00.001
TREC	CHI2	128	45	86,31 ± 9,28	48,71 ± 14,82	20,70 ± 29,84	80,27 ± 29,57	57,73 ± 1,06	00:00:00.065 ± 00:00:00.000	00:00:00.130 ± 00:00:00.001
TREC	CHI2	256	68	68,94 ± 8,46	76,20 ± 18,22	70,46 ± 33,84	31,81 ± 33,43	59,89 ± 1,89	00:00:00.104 ± 00:00:00.004	00:00:00.150 ± 00:00:00.002
TREC	CHI2	512	133	70,62 ± 10,41	78,03 ± 18,58	70,89 ± 33,28	33,55 ± 32,57	61,28 ± 1,68	00:00:00.381 ± 00:00:00.002	00:00:00.322 ± 00:00:00.002
TREC	CHI2	1024	47	70,58 ± 6,89	58,67 ± 13,92	50,95 ± 35,88	53,06 ± 34,38	57,58 ± 1,97	00:00:00.400 ± 00:00:00.003	00:00:00.279 ± 00:00:00.001
TREC	FD	8	6	75,68 ± 0,71	60,08 ± 1,39	39,98 ± 4,65	87,44 ± 1,81	65,72 ± 1,03	00:00:00.216 ± 00:00:00.002	00:00:00.175 ± 00:00:00.001
TREC	FD	16	7	78,18 ± 0,70	62,29 ± 1,13	44,92 ± 3,41	87,81 ± 1,29	68,24 ± 0,77	00:00:00.247 ± 00:00:00.002	00:00:00.182 ± 00:00:00.001
TREC	FD	32	14	77,50 ± 0,59	66,41 ± 1,26	55,81 ± 3,20	84,28 ± 1,30	70,98 ± 0,74	00:00:00.431 ± 00:00:00.007	00:00:00.274 ± 00:00:00.002
TREC	FD	64	7	71,98 ± 3,75	65,79 ± 9,63	48,95 ± 20,66	78,54 ± 13,33	66,21 ± 3,33	00:00:00.219 ± 00:00:00.002	00:00:00.165 ± 00:00:00.001
TREC	FD	128	22	68,93 ± 0,36	85,20 ± 4,29	88,45 ± 4,18	61,51 ± 1,37	76,01 ± 1,85	00:00:00.632 ± 00:00:00.002	00:00:00.332 ± 00:00:00.001
TREC	FD	256	24	77,14 ± 2,76	72,96 ± 6,78	66,15 ± 11,55	79,96 ± 7,05	74,04 ± 2,14	00:00:00.582 ± 00:00:00.004	00:00:00.344 ± 00:00:00.030
TREC	FD	512	51	69,24 ± 0,43	83,35 ± 4,69	86,38 ± 4,32	63,09 ± 1,12	75,50 ± 2,06	00:00:01.553 ± 00:00:00.048	00:00:00.579 ± 00:00:00.002
TREC	FD	1024	49	69,19 ± 0,30	87,13 ± 4,64	89,99 ± 4,33	61,55 ± 1,42	76,94 ± 1,95	00:00:01.372 ± 00:00:00.049	00:00:00.546 ± 00:00:00.003
TREC	MI	8	5	75,07 ± 0,57	58,13 ± 1,63	39,21 ± 4,57	86,64 ± 1,17	64,71 ± 1,40	00:00:00.202 ± 00:00:00.008	00:00:00.145 ± 00:00:00.002
TREC	MI	16	6	78,99 ± 0,28	61,10 ± 1,53	43,05 ± 4,60	88,57 ± 1,40	67,86 ± 1,16	00:00:00.234 ± 00:00:00.004	00:00:00.157 ± 00:00:00.001
TREC	MI	32	9	62,48 ± 0,87	81,74 ± 6,45	87,68 ± 5,29	49,29 ± 1,31	70,24 ± 2,80	00:00:00.293 ± 00:00:00.011	00:00:00.184 ± 00:00:00.001
TREC	MI	64	16	81,24 ± 6,84	71,46 ± 10,72	56,64 ± 24,00	82,18 ± 11,49	72,58 ± 4,32	00:00:00.438 ± 00:00:00.012	00:00:00.251 ± 00:00:00.001
TREC	MI	128	26	74,77 ± 0,12	76,76 ± 7,13	73,94 ± 10,80	75,81 ± 3,62	75,32 ± 3,56	00:00:00.711 ± 00:00:00.006	00:00:00.350 ± 00:00:00.002
TREC	MI	256	37	80,41 ± 4,35	73,35 ± 8,48	63,95 ± 15,75	82,66 ± 8,69	75,04 ± 2,83	00:00:00.848 ± 00:00:00.004	00:00:00.435 ± 00:00:00.001
TREC	MI	512	66	80,87 ± 2,04	70,27 ± 6,25	60,15 ± 10,79	85,52 ± 5,09	74,17 ± 2,49	00:00:01.386 ± 00:00:00.076	00:00:00.689 ± 00:00:00.003
TREC	MI	1024	204	85,48 ± 1,84	81,31 ± 6,43	77,36 ± 9,78	86,87 ± 3,49	82,75 ± 2,73	00:00:03.994 ± 00:00:00.113	00:00:01.943 ± 00:00:00.013
Unifei 2017	CHI2	8	8	79,16 ± 14,45	66,29 ± 10,86	20,09 ± 30,12	80,01 ± 30,13	59,28 ± 0,64	00:00:00.043 ± 00:00:00.000	00:00:00.351 ± 00:00:00.001
Unifei 2017	CHI2	16	10	66,48 ± 19,25	73,97 ± 11,22	50,06 ± 37,64	50,06 ± 37,65	58,37 ± 1,40	00:00:00.050 ± 00:00:00.001	00:00:00.353 ± 00:00:00.001
Unifei 2017	CHI2	32	20	76,36 ± 17,48	66,89 ± 8,97	30,10 ± 34,48	70,04 ± 34,49	58,85 ± 1,51	00:00:00.060 ± 00:00:00.002	00:00:00.359 ± 00:00:00.002
Unifei 2017	CHI2	64	48	74,35 ± 16,51	66,58 ± 8,55	30,13 ± 34,46	70,05 ± 34,48	58,47 ± 1,44	00:00:00.081 ± 00:00:00.001	00:00:00.369 ± 00:00:00.001
Unifei 2017	CHI2	128	86	80,24 ± 14,85	64,54 ± 7,92	20,28 ± 30,03	80,03 ± 30,07	59,20 ± 1,24	00:00:00.146 ± 00:00:00.012	00:00:00.395 ± 00:00:00.002
Unifei 2017	CHI2	256	34	81,68 ± 15,53	61,21 ± 2,53	20,00 ± 29,96	80,15 ± 29,92	58,65 ± 2,47	00:00:00.110 ± 00:00:00.004	00:00:00.386 ± 00:00:00.002
Unifei 2017	CHI2	512	45	74,81 ± 8,42	59,80 ± 1,15	11,59 ± 21,46	90,02 ± 21,67	57,83 ± 1,57	00:00:00.385 ± 00:00:00.008	00:00:00.525 ± 00:00:00.001
Unifei 2017	CHI2	1024	17	65,72 ± 15,66	77,98 ± 15,70	50,62 ± 37,15	53,51 ± 34,93	60,37 ± 0,75	00:00:00.648 ± 00:00:00.004	00:00:00.614 ± 00:00:00.002
Unifei 2017	FD	8	5	54,67 ± 0,61	65,88 ± 0,23	84,29 ± 1,86	30,20 ± 3,26	58,72 ± 0,47	00:00:00.654 ± 00:00:00.009	00:00:00.528 ± 00:00:00.000
Unifei 2017	FD	16	3	55,99 ± 1,15	58,02 ± 6,86	47,23 ± 18,35	64,06 ± 11,96	56,31 ± 3,59	00:00:00.359 ± 00:00:00.068	00:00:00.351 ± 00:00:00.068
Unifei 2017	FD	32	6	59,14 ± 0,76	68,90 ± 1,80	79,17 ± 3,38	45,27 ± 4,16	63,10 ± 0,06	00:00:00.685 ± 00:00:00.005	00:00:00.556 ± 00:00:00.002
Unifei 2017	FD	64	22	70,29 ± 0,75	67,92 ± 0,58	65,73 ± 1,73	72,24 ± 1,70	69,04 ± 0,09	00:00:02.145 ± 00:00:00.009	00:00:01.277 ± 00:00:00.003
Unifei 2017	FD	128	65	75,67 ± 0,75	78,05 ± 0,82	78,89 ± 1,41	74,65 ± 1,49	76,81 ± 0,07	00:00:06.320 ± 00:00:00.059	00:00:03.008 ± 00:00:00.018
Unifei 2017	FD	256	51	78,44 ± 1,76	71,71 ± 1,35	67,66 ± 3,22	81,22 ± 2,89	74,76 ± 0,06	00:00:05.058 ± 00:00:00.242	00:00:02.339 ± 00:00:00.004
Unifei 2017	FD	512	101	77,16 ± 0,95	79,45 ± 0,78	80,16 ± 1,38	76,28 ± 1,62	78,26 ± 0,13	00:00:09.468 ± 00:00:00.363	00:00:04.199 ± 00:00:00.021
Unifei 2017	FD	1024	105	75,10 ± 1,95	74,16 ± 2,24	73,19 ± 4,31	75,41 ± 4,24	74,47 ± 0,10	00:00:10.872 ± 00:00:00.655	00:00:04.242 ± 00:00:00.014
Unifei 2017	MI	8	4	57,00 ± 2,10	71,94 ± 2,69	82,34 ± 6,73	40,23 ± 9,19	62,82 ± 0,51	00:00:00.443 ± 00:00:00.001	00:00:00.412 ± 00:00:00.001
Unifei 2017	MI	16	7	73,10 ± 6,18	57,21 ± 5,36	27,29 ± 21,32	84,76 ± 17,37	60,20 ± 0,92	00:00:00.631 ± 00:00:00.003	00:00:00.521 ± 00:00:00.009
Unifei 2017	MI	32	5	64,23 ± 7,86	65,74 ± 6,86	59,30 ± 27,38	58,45 ± 20,06	61,64 ± 1,88	00:00:00.542 ± 00:00:00.003	00:00:00.447 ± 00:00:00.005
Unifei 2017	MI	64	9	62,69 ± 1,05	77,08 ± 4,61	84,01 ± 5,63	49,71 ± 5,65	68,32 ± 0,85	00:00:01.065 ± 00:00:00.021	00:00:00.637 ± 00:00:00.001
Unifei 2017	MI	128	18	78,10 ± 9,52	63,60 ± 5,65	49,52 ± 19,62	78,88 ± 14,11	67,22 ± 0,64	00:00:01.545 ± 00:00:00.001	00:00:00.958 ± 00:00:00.002
Unifei 2017	MI	256	31	76,71 ± 2,55	72,69 ± 3,90	69,26 ± 7,58	78,27 ± 5,55	74,23 ± 0,85	00:00:02.874 ± 00:00:00.012	00:00:01.482 ± 00:00:00.005
Unifei 2017	MI	512	125	82,97 ± 1,84	74,89 ± 3,08	70,71 ± 5,59	85,19 ± 3,14	78,44 ± 0,93	00:00:13.367 ± 00:00:00.154	00:00:05.031 ± 00:00:00.026
Unifei 2017	MI	1024	208	84,30 ± 1,64	78,06 ± 2,12	75,53 ± 3,62	85,77 ± 2,48	80,91 ± 0,41	00:00:22.785 ± 00:00:01.653	00:00:08.090 ± 00:00:00.028
Unifei 2018	CHI2	8	4	58,85 ± 20,22	87,62 ± 14,09	70,35 ± 34,15	30,23 ± 34,44	59,59 ± 1,09	00:00:00.051 ± 00:00:00.002	00:00:00.353 ± 00:00:00.001
Unifei 2018	CHI2	16	16	70,63 ± 21,86	78,96 ± 14,92	51,00 ± 36,93	50,66 ± 37,20	60,48 ± 1,23	00:00:00.075 ± 00:00:00.004	00:00:00.365 ± 00:00:00.001
Unifei 2018	CHI2	32	28	64,69 ± 21,27	83,34 ± 14,71	61,07 ± 35,94	41,12 ± 36,24	60,41 ± 1,19	00:00:00.114 ± 00:00:00.002	00:00:00.385 ± 00:00:00.001
Unifei 2018	CHI2	64	42	78,73 ± 18,32	71,23 ± 13,28	32,15 ± 33,46	70,66 ± 33,65	60,97 ± 0,92	00:00:00.240 ± 00:00:00.002	00:00:00.423 ± 00:00:00.001
Unifei 2018	CHI2	128	69	64,57 ± 19,70	80,52 ± 12,72	63,18 ± 33,67	42,57 ± 35,22	61,14 ± 1,75	00:00:00.778 ± 00:00:00.019	00:00:00.666 ± 00:00:00.002
Unifei 2018	CHI2	256	116	63,88 ± 19,30	81,28 ± 12,83	63,18 ± 33,63	43,28 ± 34,73	61,39 ± 1,52	00:00:00.992 ± 00:00:00.017	00:00:00.924 ± 00:00:00.012
Unifei 2018	CHI2	512	121	77,28 ± 16,37	71,57 ± 12,54	50,15 ± 29,70	65,92 ± 30,70	65,21 ± 1,05	00:00:03.867 ± 00:00:00.226	00:00:02.168 ± 00:00:00.005
Unifei 2018	CHI2	1024	152	77,31 ± 11,68	67,01 ± 10,96	44,14 ± 26,06	75,01 ± 25,54	65,26 ± 0,03	00:00:08.455 ± 00:00:00.408	00:00:03.863 ± 00:00:00.026
Unifei 2018	FD	8	5	54,38 ± 1,53	66,63 ± 2,81	85,37 ± 0,07	29,37 ± 4,22	58,90 ± 2,12	00:00:00.634 ± 00:00:00.008	00:00:00.532 ± 00:00:00.002
Unifei 2018	FD	16	3	58,52 ± 0,44	65,27 ± 4,43	70,36 ± 9,57	50,87 ± 7,48	61,24 ± 1,51	00:00:00.412 ± 00:00:00.002	00:00:00.397 ± 00:00:00.002
Unifei 2018	FD	32	8	60,21 ± 0,70	62,17 ± 0,29	62,88 ± 1,67	59,43 ± 2,29	61,17 ± 0,35	00:00:00.911 ± 00:00:00.014	00:00:00.667 ± 00:00:00.004
Unifei 2018	FD	64	11	67,70 ± 1,97	59,70 ± 4,33	41,65 ± 16,72	79,33 ± 9,54	62,01 ± 2,45	00:00:00.985 ± 00:00:00.026	00:00:00.703 ± 00:00:00.006
Unifei 2018	FD	128	24	66,91 ± 1,01	78,17 ± 0,41	82,80 ± 0,18	60,05 ± 1,91	71,98 ± 0,79	00:00:02.505 ± 00:00:00.059	00:00:01.297 ± 00:00:00.033
Unifei 2018	FD	256	60	77,33 ± 1,47	73,93 ± 1,94	70,88 ± 4,01	79,54 ± 2,78	75,42 ± 0,43	00:00:06.935 ± 00:00:00.371	00:00:02.644 ± 00:00:00.007
Unifei 2018	FD	512	79	73,14 ± 1,04	80,95 ± 0,21	83,02 ± 0,18	70,28 ± 1,58	76,85 ± 0,67	00:00:08.322 ± 00:00:00.470	00:00:03.347 ± 00:00:00.007
Unifei 2018	FD	1024	77	70,92 ± 1,34	81,01 ± 0,86	83,99 ± 1,61	66,26 ± 2,75	75,54 ± 0,41	00:00:08.091 ± 00:00:00.239	00:00:03.196 ± 00:00:00.018
Unifei 2018	MI	8	3	58,28 ± 8,75	24,92 ± 28,71	95,70 ± 4,95	21,67 ± 24,96	47,32 ± 15,76	00:00:00.461 ± 00:00:00.001	00:00:00.498 ± 00:00:00.002
Unifei 2018	MI	16	3	53,13 ± 1,03	85,39 ± 4,08	95,97 ± 2,06	18,99 ± 5,09	62,76 ± 0,28	00:00:00.448 ± 00:00:00.001	00:00:00.416 ± 00:00:00.002
Unifei 2018	MI	32	7	58,28 ± 1,48	74,86 ± 0,39	85,63 ± 2,00	40,88 ± 4,97	64,86 ± 1,06	00:00:00.730 ± 00:00:00.016	00:00:00.587 ± 00:00:00.001
Unifei 2018	MI	64	10	71,72 ± 8,52	67,23 ± 6,20	59,76 ± 20,93	69,63 ± 13,95	66,85 ± 1,43	00:00:01.123 ± 00:00:00.030	00:00:00.652 ± 00:00:00.000
Unifei 2018	MI	128	30	64,43 ± 0,71	83,13 ± 0,39	89,15 ± 0,11	52,10 ± 1,59	72,17 ± 0,66	00:00:04.105 ± 00:00:00.146	00:00:01.484 ± 00:00:00.023
Unifei 2018	MI	256	49	77,45 ± 2,33	70,83 ± 2,18	65,30 ± 5,08	80,97 ± 4,25	73,63 ± 0,18	00:00:05.109 ± 00:00:00.073	00:00:02.181 ± 00:00:00.002
Unifei 2018	MI	512	118	78,56 ± 1,69	79,14 ± 1,69	78,43 ± 2,96	78,93 ± 2,87	78,77 ± 0,09	00:00:10.942 ± 00:00:00.330	00:00:04.756 ± 00:00:00.017
Unifei 2018	MI	1024	155	83,02 ± 2,26	79,90 ± 2,41	77,90 ± 4,24	84,11 ± 3,24	81,23 ± 0,30	00:00:14.037 ± 00:00:00.911	00:00:06.054 ± 00:00:00.018

Tabela 25: Resultados do classificador SVM (polinomial).

name	method	nf	nf'	hamPrecision	spamPrecision	hamRecall	spamRecall	fMeasure	trainingTime	testingTime
Ling Spam	CHI2	8	1	83,31 ± 0,01	94,15 ± 1,09	99,93 ± 0,01	5,33 ± 0,17	66,07 ± 0,18	00:00:00.044 ± 00:00:00.000	00:00:00.048 ± 00:00:00.003
Ling Spam	CHI2	16	1	83,16 ± 0,01	96,51 ± 0,52	99,92 ± 0,01	9,93 ± 0,27	68,17 ± 0,15	00:00:00.044 ± 00:00:00.001	00:00:00.048 ± 00:00:00.002
Ling Spam	CHI2	32	1	82,87 ± 0,01	98,37 ± 0,44	99,93 ± 0,02	17,48 ± 0,25	71,25 ± 0,12	00:00:00.044 ± 00:00:00.001	00:00:00.048 ± 00:00:00.001
Ling Spam	CHI2	64	3	82,32 ± 0,02	99,30 ± 0,14	99,94 ± 0,01	28,14 ± 0,27	75,11 ± 0,10	00:00:00.045 ± 00:00:00.002	00:00:00.048 ± 00:00:00.001
Ling Spam	CHI2	128	5	81,37 ± 0,02	99,47 ± 0,09	99,92 ± 0,01	40,44 ± 0,20	79,02 ± 0,06	00:00:00.046 ± 00:00:00.002	00:00:00.050 ± 00:00:00.001
Ling Spam	CHI2	256	14	79,94 ± 0,03	99,75 ± 0,07	99,93 ± 0,02	51,51 ± 0,16	82,18 ± 0,04	00:00:00.054 ± 00:00:00.001	00:00:00.056 ± 00:00:00.001
Ling Spam	CHI2	512	14	68,53 ± 0,20	90,45 ± 0,39	97,22 ± 0,13	37,10 ± 0,59	72,80 ± 0,23	00:00:00.090 ± 00:00:00.005	00:00:00.059 ± 00:00:00.001
Ling Spam	CHI2	1024	242	83,74 ± 0,13	93,31 ± 0,19	96,53 ± 0,13	72,08 ± 0,17	86,36 ± 0,10	00:00:24.757 ± 00:00:05.576	00:00:00.712 ± 00:00:00.003
Ling Spam	FD	8	4	96,13 ± 0,26	78,49 ± 0,33	70,87 ± 0,55	97,39 ± 0,17	85,69 ± 0,21	00:00:00.062 ± 00:00:00.002	00:00:00.054 ± 00:00:00.002
Ling Spam	FD	16	4	96,22 ± 0,31	78,66 ± 0,46	71,15 ± 0,70	97,44 ± 0,21	85,84 ± 0,25	00:00:00.062 ± 00:00:00.001	00:00:00.054 ± 00:00:00.001
Ling Spam	FD	32	10	97,51 ± 0,19	88,57 ± 0,16	86,20 ± 0,24	97,98 ± 0,15	92,56 ± 0,12	00:00:00.098 ± 00:00:00.010	00:00:00.064 ± 00:00:00.001
Ling Spam	FD	64	19	97,35 ± 0,21	91,86 ± 0,16	90,54 ± 0,23	97,75 ± 0,18	94,37 ± 0,15	00:00:00.152 ± 00:00:00.008	00:00:00.080 ± 00:00:00.002
Ling Spam	FD	128	38	98,66 ± 0,16	93,99 ± 0,16	93,09 ± 0,23	98,84 ± 0,14	96,15 ± 0,14	00:00:00.438 ± 00:00:00.034	00:00:00.119 ± 00:00:00.001
Ling Spam	FD	256	58	98,75 ± 0,14	95,72 ± 0,26	95,17 ± 0,33	98,90 ± 0,12	97,13 ± 0,15	00:00:00.844 ± 00:00:00.128	00:00:00.169 ± 00:00:00.001
Ling Spam	FD	512	103	99,13 ± 0,10	97,38 ± 0,14	97,08 ± 0,14	99,22 ± 0,08	98,20 ± 0,06	00:00:02.735 ± 00:00:00.259	00:00:00.319 ± 00:00:00.001
Ling Spam	FD	1024	121	98,15 ± 0,29	98,26 ± 0,10	98,10 ± 0,11	98,31 ± 0,26	98,20 ± 0,15	00:00:03.431 ± 00:00:00.352	00:00:00.388 ± 00:00:00.002
Ling Spam	MI	8	6	93,07 ± 6,80	87,59 ± 4,01	84,59 ± 5,73	92,62 ± 7,62	89,46 ± 1,17	00:00:00.068 ± 00:00:00.002	00:00:00.058 ± 00:00:00.002
Ling Spam	MI	16	12	98,96 ± 0,08	90,00 ± 0,19	87,97 ± 0,22	99,16 ± 0,06	94,02 ± 0,08	00:00:00.108 ± 00:00:00.007	00:00:00.068 ± 00:00:00.001
Ling Spam	MI	32	19	93,74 ± 2,54	95,31 ± 1,95	94,77 ± 2,45	93,99 ± 2,51	94,45 ± 0,21	00:00:00.174 ± 00:00:00.011	00:00:00.080 ± 00:00:00.002
Ling Spam	MI	64	34	94,89 ± 0,25	98,04 ± 0,16	97,92 ± 0,17	95,17 ± 0,22	96,51 ± 0,12	00:00:00.430 ± 00:00:00.024	00:00:00.111 ± 00:00:00.001
Ling Spam	MI	128	52	95,82 ± 0,24	98,16 ± 0,12	98,03 ± 0,12	96,08 ± 0,22	97,02 ± 0,10	00:00:00.792 ± 00:00:00.056	00:00:00.153 ± 00:00:00.001
Ling Spam	MI	256	68	96,07 ± 0,25	98,63 ± 0,16	98,53 ± 0,17	96,31 ± 0,23	97,39 ± 0,10	00:00:01.403 ± 00:00:00.106	00:00:00.197 ± 00:00:00.002
Ling Spam	MI	512	139	97,68 ± 0,18	98,45 ± 0,08	98,32 ± 0,09	97,86 ± 0,16	98,08 ± 0,07	00:00:05.120 ± 00:00:00.465	00:00:00.470 ± 00:00:00.003
Ling Spam	MI	1024	157	97,14 ± 0,27	98,63 ± 0,17	98,52 ± 0,18	97,35 ± 0,23	97,91 ± 0,14	00:00:05.895 ± 00:00:01.015	00:00:00.560 ± 00:00:00.003
Spam Assassin	CHI2	8	4	67,49 ± 1,08	75,66 ± 1,02	76,32 ± 1,67	66,59 ± 2,15	71,51 ± 0,26	00:00:00.051 ± 00:00:00.001	00:00:00.048 ± 00:00:00.002
Spam Assassin	CHI2	16	7	69,18 ± 0,57	76,23 ± 0,57	76,09 ± 0,71	69,32 ± 1,06	72,71 ± 0,35	00:00:00.070 ± 00:00:00.003	00:00:00.051 ± 00:00:00.003
Spam Assassin	CHI2	32	14	70,41 ± 0,61	77,45 ± 0,64	77,27 ± 0,93	70,59 ± 1,00	73,93 ± 0,45	00:00:00.109 ± 00:00:00.010	00:00:00.054 ± 00:00:00.001
Spam Assassin	CHI2	64	20	71,01 ± 0,71	79,38 ± 0,98	79,37 ± 1,26	70,99 ± 1,10	75,19 ± 0,56	00:00:00.180 ± 00:00:00.046	00:00:00.061 ± 00:00:00.003
Spam Assassin	CHI2	128	25	73,78 ± 1,21	80,66 ± 1,35	80,33 ± 2,00	74,03 ± 2,19	77,20 ± 0,25	00:00:00.244 ± 00:00:00.030	00:00:00.062 ± 00:00:00.001
Spam Assassin	CHI2	256	29	79,60 ± 0,59	85,36 ± 0,99	84,88 ± 1,31	80,15 ± 0,95	82,50 ± 0,38	00:00:00.406 ± 00:00:00.080	00:00:00.066 ± 00:00:00.002
Spam Assassin	CHI2	512	7	61,29 ± 0,38	92,45 ± 0,46	95,99 ± 0,26	44,74 ± 0,43	73,47 ± 0,24	00:00:00.061 ± 00:00:00.003	00:00:00.049 ± 00:00:00.001
Spam Assassin	CHI2	1024	20	67,83 ± 0,65	79,88 ± 0,45	82,68 ± 0,46	63,68 ± 0,94	73,52 ± 0,41	00:00:00.135 ± 00:00:00.013	00:00:00.058 ± 00:00:00.000
Spam Assassin	FD	8	5	63,55 ± 0,70	84,60 ± 0,65	89,52 ± 0,53	52,85 ± 1,02	72,60 ± 0,35	00:00:00.057 ± 00:00:00.003	00:00:00.050 ± 00:00:00.002
Spam Assassin	FD	16	6	75,99 ± 0,65	88,83 ± 0,43	89,96 ± 0,37	73,75 ± 0,84	82,13 ± 0,30	00:00:00.066 ± 00:00:00.004	00:00:00.049 ± 00:00:00.001
Spam Assassin	FD	32	9	79,88 ± 0,72	90,07 ± 0,53	90,58 ± 0,44	78,94 ± 0,91	84,87 ± 0,47	00:00:00.086 ± 00:00:00.007	00:00:00.053 ± 00:00:00.001
Spam Assassin	FD	64	16	83,73 ± 0,40	91,91 ± 0,63	92,05 ± 0,59	83,49 ± 0,42	87,80 ± 0,28	00:00:00.137 ± 00:00:00.009	00:00:00.057 ± 00:00:00.000
Spam Assassin	FD	128	26	89,22 ± 0,75	92,69 ± 0,50	92,34 ± 0,50	89,70 ± 0,74	90,99 ± 0,39	00:00:00.261 ± 00:00:00.044	00:00:00.065 ± 00:00:00.001
Spam Assassin	FD	256	53	93,56 ± 0,55	96,40 ± 0,33	96,20 ± 0,34	93,88 ± 0,53	95,01 ± 0,25	00:00:00.773 ± 00:00:00.085	00:00:00.091 ± 00:00:00.001
Spam Assassin	FD	512	67	91,43 ± 0,52	95,41 ± 0,51	95,22 ± 0,51	91,76 ± 0,54	93,46 ± 0,34	00:00:01.232 ± 00:00:00.139	00:00:00.111 ± 00:00:00.003
Spam Assassin	FD	1024	58	82,66 ± 0,35	91,44 ± 0,45	91,65 ± 0,45	82,25 ± 0,47	87,00 ± 0,25	00:00:01.459 ± 00:00:00.295	00:00:00.096 ± 00:00:00.000
Spam Assassin	MI	8	3	72,31 ± 0,60	89,16 ± 0,46	90,97 ± 0,41	68,09 ± 0,71	80,13 ± 0,21	00:00:00.052 ± 00:00:00.001	00:00:00.047 ± 00:00:00.001
Spam Assassin	MI	16	6	81,25 ± 0,63	91,07 ± 0,45	91,38 ± 0,41	80,66 ± 0,75	86,09 ± 0,32	00:00:00.063 ± 00:00:00.002	00:00:00.050 ± 00:00:00.001
Spam Assassin	MI	32	18	86,57 ± 0,52	93,40 ± 0,61	93,38 ± 0,56	86,62 ± 0,59	89,99 ± 0,45	00:00:00.178 ± 00:00:00.030	00:00:00.059 ± 00:00:00.001
Spam Assassin	MI	64	33	87,59 ± 0,64	93,76 ± 0,57	93,67 ± 0,57	87,74 ± 0,68	90,69 ± 0,31	00:00:00.362 ± 00:00:00.042	00:00:00.073 ± 00:00:00.002
Spam Assassin	MI	128	58	93,51 ± 0,53	97,03 ± 0,40	96,90 ± 0,41	93,79 ± 0,49	95,31 ± 0,20	00:00:00.844 ± 00:00:00.089	00:00:00.103 ± 00:00:00.004
Spam Assassin	MI	256	85	93,91 ± 0,51	96,99 ± 0,32	96,83 ± 0,32	94,19 ± 0,54	95,48 ± 0,27	00:00:01.835 ± 00:00:00.091	00:00:00.135 ± 00:00:00.001
Spam Assassin	MI	512	96	93,28 ± 0,45	96,24 ± 0,47	96,04 ± 0,46	93,60 ± 0,47	94,79 ± 0,39	00:00:02.553 ± 00:00:00.279	00:00:00.151 ± 00:00:00.001
Spam Assassin	MI	1024	84	92,04 ± 0,46	95,18 ± 0,47	94,94 ± 0,46	92,42 ± 0,44	93,64 ± 0,29	00:00:01.869 ± 00:00:00.391	00:00:00.132 ± 00:00:00.001
TREC	CHI2	8	4	61,10 ± 0,00	97,36 ± 0,65	99,99 ± 0,00	0,30 ± 0,01	61,42 ± 0,10	00:00:00.045 ± 00:00:00.001	00:00:00.114 ± 00:00:00.002
TREC	CHI2	16	11	61,42 ± 0,00	90,68 ± 0,74	99,94 ± 0,01	0,98 ± 0,02	60,66 ± 0,12	00:00:00.054 ± 00:00:00.001	00:00:00.117 ± 00:00:00.002
TREC	CHI2	32	10	61,39 ± 0,00	90,23 ± 0,72	99,94 ± 0,00	0,89 ± 0,02	60,56 ± 0,12	00:00:00.055 ± 00:00:00.002	00:00:00.115 ± 00:00:00.000
TREC	CHI2	64	12	61,37 ± 0,00	93,56 ± 0,39	99,96 ± 0,00	1,02 ± 0,01	61,13 ± 0,06	00:00:00.059 ± 00:00:00.003	00:00:00.117 ± 00:00:00.002
TREC	CHI2	128	45	61,70 ± 0,01	87,94 ± 0,21	99,84 ± 0,00	1,82 ± 0,02	60,54 ± 0,04	00:00:00.833 ± 00:00:00.162	00:00:00.149 ± 00:00:00.001
TREC	CHI2	256	68	86,91 ± 0,35	39,39 ± 0,00	1,91 ± 0,02	99,55 ± 0,01	56,26 ± 0,08	00:00:02.629 ± 00:00:00.276	00:00:00.219 ± 00:00:00.003
TREC	CHI2	512	133	61,58 ± 0,01	94,21 ± 0,21	99,79 ± 0,01	5,28 ± 0,05	62,75 ± 0,05	00:00:59.570 ± 00:00:17.417	00:00:00.975 ± 00:00:00.002
TREC	CHI2	1024	47	61,88 ± 0,03	77,03 ± 0,50	98,45 ± 0,05	7,90 ± 0,05	60,23 ± 0,10	00:00:24.538 ± 00:00:05.966	00:00:00.561 ± 00:00:00.002
TREC	FD	8	6	69,33 ± 0,24	71,66 ± 0,19	71,91 ± 0,38	69,06 ± 0,45	70,49 ± 0,08	00:00:00.989 ± 00:00:00.194	00:00:00.217 ± 00:00:00.000
TREC	FD	16	7	69,48 ± 0,12	74,55 ± 0,13	76,14 ± 0,14	67,64 ± 0,18	71,95 ± 0,10	00:00:01.480 ± 00:00:00.204	00:00:00.236 ± 00:00:00.002
TREC	FD	32	14	75,59 ± 0,15	77,22 ± 0,30	76,83 ± 0,39	75,99 ± 0,25	76,41 ± 0,14	00:00:04.357 ± 00:00:00.534	00:00:00.389 ± 00:00:00.002
TREC	FD	64	7	62,99 ± 0,16	88,46 ± 0,19	93,70 ± 0,14	46,73 ± 0,29	72,87 ± 0,07	00:00:01.232 ± 00:00:00.212	00:00:00.217 ± 00:00:00.001
TREC	FD	128	22	72,19 ± 0,18	85,03 ± 0,25	87,73 ± 0,26	67,35 ± 0,27	78,07 ± 0,09	00:00:12.940 ± 00:00:01.298	00:00:00.546 ± 00:00:00.001
TREC	FD	256	24	70,11 ± 0,11	90,26 ± 0,14	93,09 ± 0,12	61,76 ± 0,18	78,78 ± 0,05	00:00:09.848 ± 00:00:01.701	00:00:00.572 ± 00:00:00.000
TREC	FD	512	51	71,54 ± 0,18	91,77 ± 0,14	94,04 ± 0,13	63,97 ± 0,34	80,31 ± 0,08	00:00:50.850 ± 00:00:16.221	00:00:01.308 ± 00:00:00.002
TREC	FD	1024	49	70,71 ± 0,15	91,07 ± 0,13	93,59 ± 0,13	62,78 ± 0,28	79,51 ± 0,08	00:01:12.034 ± 00:00:20.466	00:00:01.229 ± 00:00:00.002
TREC	MI	8	5	78,68 ± 0,18	63,68 ± 0,16	52,87 ± 0,35	85,22 ± 0,21	70,10 ± 0,09	00:00:00.639 ± 00:00:00.165	00:00:00.178 ± 00:00:00.001
TREC	MI	16	6	79,99 ± 0,12	64,52 ± 0,09	51,87 ± 0,14	87,09 ± 0,13	70,84 ± 0,06	00:00:00.809 ± 00:00:00.079	00:00:00.196 ± 00:00:00.002
TREC	MI	32	9	64,16 ± 0,12	89,71 ± 0,11	94,16 ± 0,07	49,20 ± 0,22	74,21 ± 0,07	00:00:00.817 ± 00:00:00.137	00:00:00.252 ± 00:00:00.002
TREC	MI	64	16	68,70 ± 0,18	92,53 ± 0,10	95,16 ± 0,08	58,01 ± 0,26	78,55 ± 0,11	00:00:02.635 ± 00:00:00.776	00:00:00.389 ± 00:00:00.001
TREC	MI	128	26	72,42 ± 0,13	92,01 ± 0,43	94,15 ± 0,37	65,27 ± 0,27	80,94 ± 0,15	00:00:12.814 ± 00:00:01.454	00:00:00.620 ± 00:00:00.004
TREC	MI	256	37	71,30 ± 0,15	92,16 ± 0,16	94,45 ± 0,14	63,22 ± 0,26	80,26 ± 0,09	00:00:12.838 ± 00:00:01.838	00:00:00.881 ± 00:00:00.002
TREC	MI	512	66	74,44 ± 0,24	90,48 ± 0,17	92,49 ± 0,17	69,22 ± 0,41	81,65 ± 0,13	00:01:16.864 ± 00:00:10.988	00:00:01.786 ± 00:00:00.019
TREC	MI	1024	204	81,55 ± 0,15	92,58 ± 0,17	93,36 ± 0,15	79,69 ± 0,19	86,80 ± 0,11	00:08:49.209 ± 00:01:02.402	00:00:09.555 ± 00:00:00.008
Unifei 2017	CHI2	8	8	40,96 ± 0,00	97,10 ± 0,39	100,00 ± 0,00	0,09 ± 0,00	58,02 ± 0,07	00:00:00.049 ± 00:00:00.002	00:00:00.337 ± 00:00:00.003
Unifei 2017	CHI2	16	10	40,96 ± 0,00	88,57 ± 0,70	99,98 ± 0,00	0,13 ± 0,00	56,47 ± 0,13	00:00:00.061 ± 00:00:00.003	00:00:00.342 ± 00:00:00.003
Unifei 2017	CHI2	32	20	40,95 ± 0,00	84,98 ± 0,75	99,96 ± 0,00	0,17 ± 0,00	55,78 ± 0,15	00:00:00.159 ± 00:00:00.029	00:00:00.365 ± 00:00:00.025
Unifei 2017	CHI2	64	48	40,92 ± 0,00	84,20 ± 0,72	99,94 ± 0,00	0,21 ± 0,00	55,62 ± 0,14	00:00:01.330 ± 00:00:00.395	00:00:00.388 ± 00:00:00.001
Unifei 2017	CHI2	128	86	40,85 ± 0,00	85,73 ± 0,34	99,93 ± 0,00	0,28 ± 0,00	55,93 ± 0,07	00:00:06.338 ± 00:00:01.625	00:00:00.487 ± 00:00:00.001
Unifei 2017	CHI2	256	34	40,56 ± 0,00	68,33 ± 0,37	99,46 ± 0,01	0,79 ± 0,01	52,19 ± 0,09	00:00:01.620 ± 00:00:00.377	00:00:00.421 ± 00:00:00.001
Unifei 2017	CHI2	512	45	41,34 ± 0,01	64,58 ± 0,11	96,85 ± 0,01	4,01 ± 0,01	51,67 ± 0,03	00:00:09.026 ± 00:00:01.328	00:00:00.811 ± 00:00:00.001
Unifei 2017	CHI2	1024	17	44,95 ± 0,02	98,87 ± 0,05	99,89 ± 0,01	7,18 ± 0,02	61,38 ± 0,02	00:00:04.626 ± 00:00:01.054	00:00:00.944 ± 00:00:00.001
Unifei 2017	FD	8	5	58,25 ± 0,04	62,40 ± 0,12	69,67 ± 0,25	50,20 ± 0,17	60,13 ± 0,05	00:00:02.317 ± 00:00:00.409	00:00:00.677 ± 00:00:00.008
Unifei 2017	FD	16	3	57,70 ± 0,05	75,83 ± 0,35	88,80 ± 0,23	35,06 ± 0,11	64,26 ± 0,14	00:00:01.441 ± 00:00:00.138	00:00:00.461 ± 00:00:00.001
Unifei 2017	FD	32	6	58,08 ± 0,07	73,28 ± 0,36	86,11 ± 0,25	38,01 ± 0,06	63,82 ± 0,17	00:00:04.932 ± 00:00:00.360	00:00:00.732 ± 00:00:00.001
Unifei 2017	FD	64	22	69,89 ± 0,07	79,30 ± 0,17	83,17 ± 0,17	64,28 ± 0,08	74,15 ± 0,10	00:01:19.442 ± 00:00:13.088	00:00:02.162 ± 00:00:00.001
Unifei 2017	FD	128	65	77,13 ± 0,14	86,24 ± 0,22	88,17 ± 0,21	73,94 ± 0,20	81,37 ± 0,15	00:16:06.569 ± 00:05:11.198	00:00:07.401 ± 00:00:00.012
Unifei 2017	FD	256	51	72,06 ± 0,38	81,90 ± 0,19	85,12 ± 0,15	67,09 ± 0,60	76,54 ± 0,28	00:04:14.840 ± 00:00:30.386	00:00:05.357 ± 00:00:00.007
Unifei 2017	FD	512	101	77,02 ± 0,20	86,50 ± 0,34	88,47 ± 0,31	73,67 ± 0,22	81,41 ± 0,25	00:26:17.117 ± 00:04:36.640	00:00:13.334 ± 00:00:00.012
Unifei 2017	FD	1024	105	71,70 ± 0,33	82,10 ± 0,30	85,49 ± 0,22	66,35 ± 0,47	76,41 ± 0,31	00:19:32.658 ± 00:04:04.019	00:00:14.016 ± 00:00:00.007
Unifei 2017	MI	8	4	59,92 ± 1,77	67,87 ± 2,33	72,37 ± 5,87	53,50 ± 7,90	63,40 ± 0,40	00:00:02.673 ± 00:00:00.310	00:00:00.506 ± 00:00:00.001
Unifei 2017	MI	16	7	57,15 ± 0,02	77,70 ± 0,20	89,22 ± 0,13	35,95 ± 0,05	64,92 ± 0,07	00:00:05.180 ± 00:00:01.579	00:00:00.732 ± 00:00:00.017
Unifei 2017	MI	32	5	55,78 ± 0,13	72,95 ± 0,74	86,74 ± 0,50	34,20 ± 0,06	62,36 ± 0,33	00:00:02.081 ± 00:00:00.197	00:00:00.579 ± 00:00:00.001
Unifei 2017	MI	64	9	60,80 ± 0,10	84,83 ± 0,47	92,80 ± 0,28	40,22 ± 0,21	69,52 ± 0,19	00:00:04.448 ± 00:00:00.714	00:00:00.922 ± 00:00:00.002
Unifei 2017	MI	128	18	64,01 ± 0,14	81,41 ± 0,16	88,50 ± 0,09	50,30 ± 0,29	71,02 ± 0,15	00:00:11.205 ± 00:00:01.864	00:00:01.636 ± 00:00:00.004
Unifei 2017	MI	256	31	70,36 ± 0,15	83,82 ± 0,27	87,77 ± 0,23	63,13 ± 0,26	76,26 ± 0,17	00:00:53.722 ± 00:00:10.160	00:00:02.916 ± 00:00:00.003
Unifei 2017	MI	512	125	78,26 ± 0,15	85,68 ± 0,12	87,29 ± 0,13	75,82 ± 0,19	81,76 ± 0,12	00:39:22.091 ± 00:07:53.810	00:00:18.209 ± 00:00:00.028
Unifei 2017	MI	1024	208	79,30 ± 0,07	87,14 ± 0,16	88,61 ± 0,16	76,95 ± 0,07	83,00 ± 0,09	01:32:39.551 ± 00:28:47.995	00:00:40.832 ± 00:00:00.019
Unifei 2018	CHI2	8	4	99,78 ± 0,03	59,07 ± 0,00	1,16 ± 0,01	100,00 ± 0,00	61,80 ± 0,01	00:00:00.057 ± 00:00:00.002	00:00:00.339 ± 00:00:00.002
Unifei 2018	CHI2	16	16	99,60 ± 0,06	59,17 ± 0,00	2,03 ± 0,01	99,99 ± 0,00	62,11 ± 0,01	00:00:00.622 ± 00:00:00.176	00:00:00.370 ± 00:00:00.007
Unifei 2018	CHI2	32	28	99,26 ± 0,12	59,45 ± 0,00	2,70 ± 0,01	99,99 ± 0,00	62,35 ± 0,02	00:00:01.600 ± 00:00:00.295	00:00:00.429 ± 00:00:00.024
Unifei 2018	CHI2	64	42	41,65 ± 0,00	98,48 ± 0,11	99,94 ± 0,00	2,57 ± 0,01	59,20 ± 0,02	00:00:17.901 ± 00:00:02.412	00:00:00.528 ± 00:00:00.005
Unifei 2018	CHI2	128	69	98,75 ± 0,03	59,95 ± 0,00	8,66 ± 0,02	99,92 ± 0,00	64,47 ± 0,01	00:01:13.126 ± 00:00:12.621	00:00:01.290 ± 00:00:00.006
Unifei 2018	CHI2	256	116	97,55 ± 0,04	60,51 ± 0,00	8,71 ± 0,02	99,84 ± 0,00	64,36 ± 0,01	00:04:15.076 ± 00:00:33.688	00:00:02.557 ± 00:00:00.009
Unifei 2018	CHI2	512	121	50,84 ± 0,01	94,99 ± 0,17	99,03 ± 0,04	16,07 ± 0,03	64,33 ± 0,04	00:13:04.199 ± 00:02:27.330	00:00:07.561 ± 00:00:00.018
Unifei 2018	CHI2	1024	152	53,77 ± 0,03	88,80 ± 0,08	96,71 ± 0,03	23,86 ± 0,03	65,33 ± 0,03	00:43:02.330 ± 00:03:56.289	00:00:16.233 ± 00:00:00.019
Unifei 2018	FD	8	5	60,34 ± 0,06	62,11 ± 0,17	63,35 ± 0,34	59,06 ± 0,22	61,21 ± 0,08	00:00:02.466 ± 00:00:00.119	00:00:00.689 ± 00:00:00.022
Unifei 2018	FD	16	3	58,42 ± 0,06	76,35 ± 0,08	87,76 ± 0,04	38,76 ± 0,07	65,26 ± 0,05	00:00:01.856 ± 00:00:00.180	00:00:00.476 ± 00:00:00.002
Unifei 2018	FD	32	8	59,53 ± 0,10	76,05 ± 0,87	86,09 ± 0,74	42,96 ± 0,28	66,12 ± 0,36	00:00:14.671 ± 00:00:01.196	00:00:00.910 ± 00:00:00.003
Unifei 2018	FD	64	11	61,05 ± 0,07	73,10 ± 0,12	81,30 ± 0,10	49,48 ± 0,07	66,22 ± 0,08	00:00:28.678 ± 00:00:03.671	00:00:01.060 ± 00:00:00.001
Unifei 2018	FD	128	24	68,56 ± 0,36	82,27 ± 0,33	86,43 ± 0,19	61,39 ± 0,57	74,65 ± 0,35	00:03:01.086 ± 00:00:27.565	00:00:02.247 ± 00:00:00.005
Unifei 2018	FD	256	60	74,24 ± 0,14	85,62 ± 0,11	87,88 ± 0,10	70,30 ± 0,22	79,51 ± 0,08	00:06:35.527 ± 00:01:54.698	00:00:06.490 ± 00:00:00.014
Unifei 2018	FD	512	79	77,34 ± 1,15	83,07 ± 0,32	84,10 ± 0,45	75,96 ± 1,68	80,12 ± 0,68	00:24:55.043 ± 00:05:16.464	00:00:09.352 ± 00:00:00.010
Unifei 2018	FD	1024	77	71,57 ± 0,21	85,81 ± 0,14	88,89 ± 0,07	65,56 ± 0,34	77,95 ± 0,18	00:14:44.317 ± 00:02:36.098	00:00:08.962 ± 00:00:00.013
Unifei 2018	MI	8	3	85,99 ± 0,13	63,68 ± 0,02	49,04 ± 0,08	91,79 ± 0,09	72,56 ± 0,04	00:00:00.762 ± 00:00:00.002	00:00:00.596 ± 00:00:00.001
Unifei 2018	MI	16	3	74,18 ± 0,24	72,35 ± 0,11	69,13 ± 0,10	77,05 ± 0,27	73,17 ± 0,16	00:00:01.419 ± 00:00:00.163	00:00:00.493 ± 00:00:00.002
Unifei 2018	MI	32	7	79,43 ± 0,14	74,69 ± 0,47	70,90 ± 0,70	82,39 ± 0,08	76,85 ± 0,32	00:00:04.391 ± 00:00:00.823	00:00:00.796 ± 00:00:00.001
Unifei 2018	MI	64	10	61,85 ± 0,10	80,77 ± 0,08	88,82 ± 0,06	46,16 ± 0,23	69,35 ± 0,09	00:00:05.061 ± 00:00:00.726	00:00:00.978 ± 00:00:00.002
Unifei 2018	MI	128	30	66,41 ± 0,20	83,15 ± 0,17	88,21 ± 0,17	56,60 ± 0,39	73,57 ± 0,15	00:02:04.046 ± 00:00:26.463	00:00:02.841 ± 00:00:00.022
Unifei 2018	MI	256	49	69,92 ± 0,11	80,70 ± 0,22	84,13 ± 0,21	64,70 ± 0,16	74,86 ± 0,13	00:07:48.634 ± 00:02:05.913	00:00:05.000 ± 00:00:00.011
Unifei 2018	MI	512	118	75,32 ± 0,16	86,23 ± 0,24	88,23 ± 0,21	71,82 ± 0,19	80,40 ± 0,19	00:56:38.409 ± 00:23:15.511	00:00:16.621 ± 00:00:00.017
Unifei 2018	MI	1024	155	78,35 ± 0,21	86,54 ± 0,17	87,83 ± 0,16	76,34 ± 0,27	82,27 ± 0,17	00:58:43.293 ± 00:12:58.240	00:00:25.196 ± 00:00:00.013

Tabela 26: Detalhes do *hardware* utilizado nos experimentos.

Processador	Marca	Intel
	Modelo	Intel(R) Core(TM) i5-4570
	Frequência mínima	3,20 GHz
	Frequência máxima	3,60 GHz
	Cache	6 MB
	Soquete	FCLGA1150
Memória	Tamanho	32 GB
	Tipo	DDR3-1333
Sistema Operacional	Versão	18.3
	Codinome	sylvia
	Edição	Cinnamon 64-bit
	Descrição	Linux Mint 18.3 Sylvia

Tabela 27: Detalhes do *software* utilizado nos experimentos.

Java	openjdk version “1.8.0_162”
	OpenJDK Runtime Environment (build 1.8.0_162-8u162-b12-0ubuntu0.16.04.2-b12)
	OpenJDK 64-Bit Server VM (build 25.162-b12, mixed mode)
libatlas3-base	3.10.2-9
libopenblas-base	0.2.18-ubuntu1
WEKA	Stable 3.8.1
LIBLINEAR	2.20
LIBLINEAR-poly2	2.01

Referências

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: *Mining text data*. Boston, MA: Springer, 2012. p. 163–222.

AIZERMAN, M. A.; BRAVERMAN, E. M.; ROZONOÉR, L. I. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, v. 25, p. 821–837, 1964.

ALI, A. Comparison study between IPV4 & IPV6. *International Journal of Computer Science Issues,*, Citeseer, v. 9, n. 3, p. 314–317, 2012. ISSN 16940784.

ANDROUTSOPOULOS, I. et al. An evaluation of naive bayesian anti-spam filtering. *arXiv preprint cs/0006013*, Barcelona, p. 9–17, 2000.

ANDROUTSOPOULOS, I. et al. *Ling-Spam Corpus*. 2000. Disponível em: <https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/lingspam{_}public.tar>.

ANDROUTSOPOULOS, I.; PALIOURAS, G.; MICHELAKIS, E. *PU1, PU2, PU3, and PUA corpora*. "DEMOKRITOS", National Center for Scientific Research, 2003. Disponível em: <https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/>.

ANGELINE, P. J.; SAUNDERS, G. M.; POLLACK, J. B. An evolutionary algorithm that constructs recurrent neural networks. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, v. 5, n. 1, p. 54–65, 1994. ISSN 1045-9227.

ARAGÃO, M. V. et al. Factorial design analysis applied to the performance of SMS anti-spam filtering systems. *Expert Systems with Applications*, v. 64, p. 589–604, 2016. ISSN 09574174.

BARIGOU, F.; BELDJILALI, B.; ATMANI, B. Using cellular automata for improving knn based spam filtering. *International Arab Journal of Information Technology (IAJIT)*, v. 11, n. 4, p. 345–353, 2014.

BEN-HUR, A. et al. Support vector clustering. *Journal of machine learning research*, v. 2, n. Dec, p. 125–137, 2001.

BERGER, H.; MERKL, D. A comparison of support vector machines and self-organizing maps for e-mail categorization. In: *Proceedings of AusDM 2005, the 4th Australasian Data Mining Conference*. Sydney, Australia: [s.n.], 2005. p. 189–204. ISBN 1863657169.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 1. ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 978-0-387-31073-2.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. *Proceedings of the 5th ACM Workshop on Computational Learning Theory*. Pittsburgh, PA, USA, 1992. p. 144–152.

BREIMAN, L. et al. *Classification and regression trees*. 1. ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 1984. 358 p. ISSN 19424787. ISBN 978-0412048418.

BREUNIG, M. M. et al. LOF: Identifying Density-Based Local Outliers. In: ACM. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA, 2000. v. 29, n. 2, p. 1–12. ISBN 1581132182. ISSN 01635808.

BROOMHEAD, D. S.; LOWE, D. *Radial basis functions, multi-variable functional interpolation and adaptive networks*. Malvern, Worcs, UK, 1988.

CAMPS-VALLS, G. et al. Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 3, n. 1, p. 93–97, 2006. ISSN 1545598X.

CARPINTEIRO, O. A. S. et al. A neural model in anti-spam systems. In: SPRINGER. *International Conference on Artificial Neural Networks*. Berlin, Heidelberg, Germany, 2006. p. 847–855.

CHANG, C.-C.; LIN, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, Acm, v. 2, n. 3, p. 1–27, 2011. ISSN 21576904.

CHANG, Y.-W. et al. Training and Testing Low-degree Polynomial Data Mappings via Linear SVM. *Journal of Machine Learning Research*, v. 11, n. Apr, p. 1471–1490, 2010. ISSN 1532-4435.

CHAPELLE, O. et al. Choosing mulitple parameters for support vector machines. *Machine learning*, Springer, v. 46, n. 1, p. 131–159, 2002.

COHEN, W. W. Enron Email Dataset. 2008. Disponível em: <https://www.cs.cmu.edu/{~}./enro>.

CORMACK, G. V.; LYNAM, T. R. *2005 TREC Public Spam Corpus*. 2005. Disponível em: <https://plg.uwaterloo.ca/{~}gvcormac/treccorpu>.

CORMACK, G. V.; LYNAM, T. R. *2006 TREC Public Spam Corpora*. 2006. Disponível em: <https://plg.uwaterloo.ca/{~}gvcormac/treccorpus0>.

CORMACK, G. V.; LYNAM, T. R. *2007 TREC Public Spam Corpus*. 2007. Disponível em: <http://plg.uwaterloo.ca/{~}gvcormac/treccorpus0>.

CORMACK, G. V.; OTHERS. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, Now Publishers, Inc., v. 1, n. 4, p. 335–455, 2008. ISSN 1554-0669.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. ISSN 15730565.

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. 2. ed. New York: John Wiley & Sons, 1991. 563 p.

COVÕES, T. F.; HRUSCHKA, E. R. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, Elsevier, v. 181, n. 18, p. 3766–3782, 2011. ISSN 00200255.

CYGANIAK, R.; WOOD, D.; LANTHALER, M. RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation 25 February 2014*, World Wide Web Consortium, n. February, p. 263–270, 2014. ISSN 13514180.

DEB, K. et al. A Fast and Elitist Multiobjective Genetic Algorithm: NGSA-II. *IEEE Transactions on Evolutionary Computing*, IEEE, v. 6, n. 2, p. 182–197, 2002. ISSN 1089778X.

DEERWESTER, S. et al. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, American Documentation Institute, v. 41, n. 6, p. 391–407, 1990. ISSN 10974571.

DENG, L.; YU, D. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, Now Publishers, Inc., v. 7, n. 3-4, p. 197–387, 2014. ISSN 1932-8346.

DONOHO, D. L. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. *American Math. Society Lecture-Math Challenges of the 21st Century*, Citeseer, v. 1, p. 1–33, 2000.

DREDZE, M.; GEVARYAHU, R.; ELIAS-BACHRACH, A. Image Spam Dataset. 2007. Disponível em: <http://www.cs.jhu.edu/{~}mdredze/datasets/image{_}>.

DREDZE, M.; GEVARYAHU, R.; ELIAS-BACHRACH, A. Learning Fast Classifiers for Image Spam. In: *Conference on Email and Anti-Spam (CEAS) 2007*. Berlin, Germany: [s.n.], 2007.

DRUCKER, H. et al. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, IEEE, v. 10, n. 5, p. 1048–1054, 1999.

ESPOSITO, F.; MALERBA, D.; SEMERARO, G. Decision Tree Pruning as a Search in the State Space. In: SPRINGER. *Machine Learning: ECML-93, European Conference on Machine Learning, Proceedings*. Berlin, Heidelberg, Germany, 1993. v. 667, p. 165–184. ISBN 9783540566021. ISSN 16113349.

FAN, R.-E. et al. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, v. 9, n. Aug, p. 1871–1874, 2008. ISSN 15324435.

FERREIRA, I. C. Estudo e Desenvolvimento de um Sistema Anti-Spam. 2018.

FORMAN; GEORGE. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, v. 3, n. Mar, p. 1289–1305, 2003. ISSN 1532-4435.

FOUNDATION, T. A. S. *Apache SpamAssassin Project*. 2015. Disponível em: <http://spamassassin.apache.org/>.

FREUND, Y. Boosting a weak learning algorithm by majority. *Information and computation*, Elsevier, v. 121, n. 2, p. 256–285, 1995.

FREUND, Y.; MASON, L. The alternating decision tree learning algorithm. In: *Proceeding of the Sixteenth International Conference on Machine Learning*. San Francisco, CA, USA: [s.n.], 1999. v. 99, p. 10. ISBN 9780874216561. ISSN 14602431.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a New Boosting Algorithm. In: *Proceedings of the 13th International Conference on Machine Learning*. Bari, Italy: [s.n.], 1996. v. 96, p. 148–156.

FRIEDMAN, J. H. Greedy function aproximation: A gradient boosting machine. *The Annals of Statistics*, JSTOR, v. 29, n. 5, p. 1189–1232, 2001.

F.R.S., K. P. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901.

GINI, C. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, Springer, v. 46, n. 1-3, p. 389–422, 2002. ISSN 08856125.

HALL, M. et al. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.

HALL, M. A. Correlation-based feature selection of discrete and numeric class machine learning. University of Waikato, Department of Computer Science, 2000.

HAYKIN, S. S. *Redes neurais: princípios e prática*. 2. ed. Porto Alegre, RS, Brazil: Bookman Companhia Editora, 2001. 900 p. ISBN 8573077182.

HEBB, D. O. *The Organization of Behavior; A Neuropsychological Theory*. New York, New York, USA: Wiley & Sons, 1949. 378 p.

HINTON, G. E.; ROWEIS, S. T. Stochastic Neighbor Embedding. In: *Advances in Neural Information Processing Systems 15*. Vancouver, Canada: [s.n.], 2002. p. 833–840. ISBN 0262025507. ISSN 10495258.

HOFMANN, T. Probabilistic Latent Semantic Indexing. In: ACM. *Sigir'99*. New York, NY, USA, 1999. v. 51, n. 2, p. 50–57. ISBN 1581130961. ISSN 15206882.

HOPKINS, M. et al. *Spam base dataset*. 1999. Disponível em: <https://archive.ics.uci.edu/ml/datasets/spambase>.

HUGHES, G. E. On the Mean Accuracy of Statistical Pattern Recognition Accuracy. *IEEE Transactions on Information Theory*, IEEE, IT-14, n. 1, p. 55–63, 1968.

IDRIS, I. et al. A combined negative selection algorithm–particle swarm optimization for an email spam detection system. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 39, p. 33–44, 2015.

J, K.; C, E. R. Particle swarm optimization. In: *Proceedings of IEEE International Conference on Neural Networks*. Perth, WA, Australia: Springer, 1995. p. 1942–1948. ISBN 9781612840529.

JIMENEZ, F. et al. An evolutionary algorithm for constrained multi-objective optimization. In: IEEE. *Proceedings of the 2002 Congress on Evolutionary Computation*. Honolulu, HI, USA, 2002. v. 2, p. 1133–1138.

JUNG, J.; SIT, E. An empirical study of spam traffic and the use of DNS black lists. In: ACM. *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*. Taormina, Sicily, Italy, 2004. p. 370–375. ISBN 1-58113-821-0.

KAYA, Y.; ERTUğRUL, Ö. F. A novel approach for spam email detection based on shifted binary patterns. *Security and Communication Networks*, Wiley Online Library, v. 9, n. 10, p. 1216–1225, 2016. ISSN 19390122.

KIRK, D. NVIDIA CUDA Software and GPU Parallel Computing Architecture. In: *Proceedings of the 6th International Symposium on Memory Management*. Montreal, QC, Canada: [s.n.], 2007. v. 7, p. 103–104. ISBN 9781595938930.

KOHONEN, T. Learning Vector Quantization. In: *Self-Organizing Maps*. Berlin, Heidelberg, Germany: Springer, 1995. p. 175–189. ISBN 3642976123.

KUMARESAN, T.; SARAVANAKUMAR, S.; BALAMURUGAN, R. Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine. *Cluster Computing*, Springer, p. 1–14, 2017.

LIN, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, IEEE, v. 37, n. 1, p. 145–151, 1991.

LIU, H.; SETIONO, R. A probabilistic approach to feature selection - a filter solution. In: UNIVERSITA' DI BARI, UNIVERSITA' DI TORINO. *Proc 13th International Conference on Machine Learning*. Bari, Italy, 1996. v. 96, p. 319–327.

LODHI, H. et al. Text Classification using String Kernels. *Journal of Machine Learning Research*, v. 2, n. Feb, p. 419–444, 2002. ISSN 0003-6951.

LÓPEZ-IBÁÑEZ, M.; PAQUETE, L.; STÜTZLE, T. Exploratory analysis of stochastic local search algorithms in biobjective optimization. In: *Experimental Methods for the Analysis of Optimization Algorithms*. Berlin, Heidelberg, Germany: Springer, 2010. p. 209–222. ISBN 978-3-642-02537-2.

MAATEN, L. van der; HINTON, G. Visualizing data using t-SNE. *Journal of machine learning research*, v. 9, n. Nov, p. 2579–2605, 2008.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: STATISTICAL LABORATORY OF THE UNIVERSITY OF CALIFORNIA, BERKELEY. *Proceedings of the 5th Berkeley Symposium on Math., Stat. and Prob.* Oakland, CA, USA: University of California Press, 1967. v. 1, n. 14, p. 281–296.

McAfee Inc.; ICF International. The Carbon Footprint of Email Spam Report. *McAfee Resource Center*, p. 1–28, 2009.

MCCALLUM, A.; NIGAM, K. A Comparison of Event Models for Naive Bayes Text Classification. In: JUST RESEARCH AND CARNEGIE MELLON

UNIVERSITY SCHOOL OF COMPUTER SCIENCE. *AAAI/ICML-98 Workshop on Learning for Text Categorization*. Madison, WI, USA, 1998. v. 752, n. 1, p. 41–48. ISBN 0897915240. ISSN 0343-6993.

MCCORMICK, C. *AdaBoost Tutorial*. 2013. 1–9 p. Disponível em: <http://mccormickml.com/2013/12/13/adaboost-tutorial/>.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. ISSN 0007-4985.

MCGONAGLE, J. *Naive Bayes Classifier*. 2018. 1–4 p. Disponível em: <https://brilliant.org/wiki/naive-bayes-classifier/>.

MÉNDEZ, J. R. et al. Grindstone4Spam: An optimization toolkit for boosting e-mail classification. *Journal of Systems and Software*, Elsevier, v. 85, n. 12, p. 2909–2920, 2012. ISSN 01641212.

MEYER, T. A.; WHATELEY, B. SpamBayes: Effective open-source, Bayesian based, email classification system. In: *Conference on Email and Anti-Spam (CEAS) 2004*. Mountain View, CA, USA: [s.n.], 2004.

MINGERS, J. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, Springer, v. 4, n. 2, p. 227–243, 1989.

MINGERS, J. An empirical comparison of selection measures for decision-tree induction. *Machine learning*, Springer, v. 3, n. 4, p. 319–342, 1989.

NEYMAN, J. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society, v. 236, n. 767, p. 333–380, 1937.

PAWLAK, Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, Netherlands: Springer Science & Business Media, 1991. v. 9.

PÉREZ-DÍAZ, N. et al. Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification. *Applied Soft Computing Journal*, Elsevier, v. 12, n. 11, p. 3671–3682, 2012. ISSN 15684946.

POGGIO, T.; GIROSI, F. *A theory of networks for approximation and learning*. Cambridge, MA, USA, 1989. 1–65 p.

POWELL, M. J. D. Radial basis functions for multivariable interpolation: a review. *Algorithms for approximation*, Clarendon Press, p. 143–167, 1987.

POWERS, D. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. ISSN 2229-3981.

QUINLAN, J. R. Induction of Decision Trees. *Machine Learning*, Springer, v. 1, n. 1, p. 81–106, 1986. ISSN 1573-0565.

QUINLAN, J. R. Simplifying decision trees. *International journal of man-machine studies*, Elsevier, v. 27, n. 3, p. 221–234, 1987.

QUINLAN, J. R. *C 4.5: Programs for machine learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993. 1–302 p. ISBN 1558602380.

RADICATI, S. *Email statistics report, 2018-2022*. Palo Alto, CA, USA, 2018. 1–3 p. Disponível em: <https://www.radicati.com/?p=15185>.

ROKACH, L.; MAIMON, O. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 35, n. 4, p. 476–487, 2005. ISSN 1094-6977.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. *Learning internal representations by error propagation*. La Jolla, CA, USA, 1985. 318–362 p.

RUSSELL, S.; NORVIG, P. *Inteligencia Artificial*. Rio de Janeiro, RJ, Brazil: Elsevier Editora, 2013. 1–1016 p. ISBN 978-8535237016.

SAHAMI, M. Learning Limited Dependence Bayesian Classifiers. In: *Second International Conference on Knowledge Discovery and Data Mining*. Portland, OR, USA: [s.n.], 1996. v. 96, n. 1, p. 91–94. ISBN 1-57735-004-9.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, n. 3, p. 210–229, 1959.

SCHAPIRE, R. E. The strength of weak learnability. *Machine learning*, Springer, v. 5, n. 2, p. 197–227, 1990.

SHAMS, R.; MERCER, R. E. Supervised classification of spam emails with natural language stylometry. *Neural Computing and Applications*, Springer, v. 27, n. 8, p. 2315–2331, 2016.

SIEBEL, N. N. T.; SOMMER, G. Evolutionary reinforcement learning of artificial neural networks. *International Journal of Hybrid Intelligent Systems*, IOS Press, v. 4, n. 3, p. 171–183, 2007. ISSN 1448-5869.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas. *São Paulo: Artliber*, p. 31–55, 2010.

SIRISANYALAK, B.; SORNIL, O. An artificial immunity-based spam detection system. In: IEEE. *2007 IEEE Congress on Evolutionary Computation, CEC 2007*. Singapore, 2007. p. 3392–3398. ISBN 1424413400.

SNYDER, J. *Comparing Industry-Leading Anti-Spam Services: Head-to-Head Testing Results*. Tucson, AZ, USA, 2015. 1–6 p.

SPERBER, D. et al. Epistemic vigilance. *Mind and Language*, Wiley Online Library, v. 25, n. 4, p. 359–393, 2010. ISSN 02681064.

SPRUYT, V. *The Curse of Dimensionality in classification*. 2014. 1–16 p. Disponível em: <http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>.

STANLEY, K. O.; MIIKKULAINEN, R. Evolving neural networks through augmenting topologies. *Evolutionary computation*, MIT Press, v. 10, n. 2, p. 99–127, 2002. ISSN 1063-6560.

Symantec Corporation. *Internet Security Threat Report 8*. Cupertino, CA, USA, 2005. 1–106 p. Disponível em: <https://www.symantec.com/content/dam/symantec/docs/security-center/archives/istr-05-sept-en.pdf>.

Symantec Corporation. *Internet Security Threat Report 17*. Mountain View, CA, USA, 2012. v. 17, 1–52 p. Disponível em: <https://www.symantec.com/content/dam/symantec/docs/security-center/archives/istr-12-april-volume-17-en.pdf>.

Symantec Corporation. *Internet Security Threat Report 23*. Mountain View, CA, USA, 2018. v. 23, 1–89 p. Disponível em: <https://www.symantec.com/security-center/threat-report>.

TAO, B. *CSDMC2010 SPAM corpus*. International Conference on Neural Information Processing, 2010. Disponível em: <http://csmining.org/index.php/spam-email-datasets-.html>.

TENENBAUM, J. B.; SILVA, V. de; LANGFORD, J. C. A Gobal Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000.

The Apache Software Foundation. *SpamAssassin Public Mail Corpus*. 2005. Disponível em: <https://spamassassin.apache.org/publiccorpus/>.

The Spamhaus Project. *The Spamhaus Block List*. 2018. 1–1 p. Disponível em: <https://www.spamhaus.org/sbl/>.

The Spamhaus Project. *Understanding DNSBL Filtering*. 2018. 1–2 p. Disponível em: <https://www.spamhaus.org/whitepapers/dnsbl{_}functi>.

VAPNIK, V. N.; LERNER, A. Pattern recognition using generalized portrait method. *Automation and Remote Control*, v. 24, p. 774–780, 1963.

WEBB, G. I.; BOUGHTON, J. R.; WANG, Z. Not so naive Bayes: aggregating one-dependence estimators. *Machine learning*, Springer, v. 58, n. 1, p. 5–24, 2005.

WETTSCHERECK, D.; DIETTERICH, T. Improving the Performance of Radial Basis Function Networks by Learning Center Locations. In: *Advances in Neural Information Processing Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1992. v. 4, p. 1133–1140.

WHITE, B. W.; ROSENBLATT, F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *The American Journal of Psychology*, v. 76, n. 4, p. 705, 1963. ISSN 00029556.

WIDROW, B.; HOFF, M. E. *Adaptive switching circuits*. Stanford, CA, USA, 1960.

YANG, C. et al. An Anti-Spam Filter Based on One-Class IB Method in Small Training Sets. *International Arab Journal of Information Technology (IAJIT)*, v. 13, n. 6, p. 677–685, 2016. ISSN 1683-3198.

YANG, J.; ESTIVILL-CASTRO, V.; CHALUP, S. K. Support vector clustering through proximity graph modelling. In: IEEE. *Proceedings of the 9th International Conference on Neural Information Processing*. Singapore, 2002. v. 2, p. 898–903. ISBN 9810475241.

YANG, X.-S.; DEB, S. Cuckoo Search via Levy Flights. In: IEEE. *Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on*. Coimbatore, India, 2010. p. 210–214. ISBN 978-1-4244-5053-4. ISSN 0941-0643.

YANG, Y.; JOACHIMS, T. Text categorization. *Scholarpedia*, v. 3, n. 5, p. 4242, 2008. ISSN 1941-6016. Disponível em: <http://www.scholarpedia.org/article/Text{_}categorizat>.

YANG, Y.; PEDERSEN, J. A Comparative Study on Feature Selection in Text Categorization. In: *Proceedings of ICML-97, 14th International Conference on Machine Learning*. Lille, France: [s.n.], 1997. v. 97, p. 412–420.

YEVSEYEVA, I. et al. Optimising anti-spam filters with evolutionary algorithms. *Expert Systems with Applications*, Elsevier, v. 40, n. 10, p. 4010–4021, 2013. ISSN 09574174.

YOUN, S. SPONGY (SPam ONtoloGY): email classification using two-level dynamic ontology. *The Scientific World Journal*, Hindawi Publishing Corporation, v. 2014, 2014. ISSN 1537744X.

ZHANG, H. et al. IPGroupRep: A novel reputation based system for anti-spam. *UIC-ATC 2009 - Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, p. 513–518, 2009.

ZITZLER, E.; LAUMANNS, M.; THIELE, L. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, p. 95–100, 2001. ISSN 03772217.