Mineração de Textos para Gestão de Clientes em Empresas de Telecomunicações
Alternative Title: Text Mining for Client Management in Telecom Companies

Dildre G. Vasques
Universidade de São Paulo - ICMC
Av. Trabalhador São-Carlense, 400
São Carlos, SP, Brasil
dildre.vasques@usp.br

Leonardo Comelli
Universidade de São Paulo - ICMC
Av. Trabalhador São-Carlense, 400
São Carlos, SP, Brasil
leonardo.comelli@gmail.com

Caetano M. Ranieri
Universidade de São Paulo - ICMC
Av. Trabalhador São-Carlense, 400
São Carlos, SP, Brasil
cmranieri@usp.br

Solange O. Rezende
Universidade de São Paulo - ICMC
Av. Trabalhador São-Carlense, 400
São Carlos, SP, Brasil
solange@icmc.usp.br

RESUMO
Em um mercado competitivo, adotar boas estratégias de negócio é um requisito fundamental para satisfazer e fidelizar clientes. Este trabalho teve como objetivo identificar os perfis de usuários que se mantêm fiéis a uma operadora de telecomunicações e os perfis daqueles que a abandonam. Consideram-se os relatórios de ocorrências de atendimento por telefone, registradas em uma base de dados privada, cedida para esta pesquisa. Utilizou-se Mineração de Textos para classificação das ocorrências, a fim de prever cancelamentos, e também extração de regras de associação, a fim de entender os motivos que os levam a ocorrer. Nas condições experimentadas, foi possível prever cancelamentos com acurácia de até 97,02%. Além disso, foram extraídos os atributos mais representativos para cada classe, a fim de fornecer o arcabouço para otimizar a tomada de decisões estratégicas.

Palavras-Chave
Mineração de Textos, Gestão de Clientes, Gestão do Conhecimento, Serviços de Telecomunicações.

ABSTRACT
In a competitive market, adopting suitable business strategies is essential to satisfice and make clients loyal. This research aimed to identify profiles of clients that kept loyal to a telecom company and users that cancelled the service. We have considered reports from occurrences of the phone attendance service, registered to a private database, granted for this research. We have applied text mining for classification of occurrences, to avoid cancellation, and extraction of association rules, to understand the reasons that lead users to cancel the service. In the experienced conditions, the algorithms could predict cancellations with 97,02% accuracy. Besides, the most representative attributes for each class were extracted, providing a framework for strategic decision-making.

SBSI 2017, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

CCS Concepts
• Information systems → Data mining
• Applied computing → Business intelligence

Keywords
Text Mining, Client Management, Knowledge Management, Telecom Services.

1. INTRODUÇÃO
As estratégias de inovação de mercado estão vinculadas às exigências dos consumidores [1] e visam sua satisfação com relação à qualidade dos produtos e dos serviços prestados [2, 3]. A exploração de bases de dados organizacionais para a extração de informações de clientes pode beneficiar os processos de negócio e a criação de novos projetos [4]. Ao analisarem os dados dos clientes, os tomadores de decisão lidam com informações que podem auxiliar na avaliação do desempenho do negócio e na geração de novos conhecimentos que servirão de base para otimizar as suas decisões [5]. Desse modo, os dados sobre os clientes se transformam em Inteligência de Negócios (do inglês *Business Intelligence*) e oferecem suporte à gestão de negócios [6].

A abordagem denominada Gestão de Relacionamentos com o Cliente, originária do termo em inglês *Customer Relationship Managment* (CRM), coloca o cliente no centro dos processos de negócio. A CRM é apoiada por um conjunto de ferramentas que automatizam as funções de contato com os clientes [7] e facilitam a recuperação dos dados e a descoberta de conhecimento a respeito desses clientes. No entanto, construir conhecimento a partir das informações relativas aos clientes por meio de dados não é uma tarefa trivial. Assim, os recursos da Inteligência Artificial (IA) surgem como aliados ao tratamento desses dados e

informações. A Descoberta de Conhecimento em Bases de Dados [8], definida como um processo que envolve seleção, pré-processamento, transformação e análise de dados, contribui para a interpretação dos resultados [9]. A aplicação de processos de Mineração de Dados (*Data Mining*) e de Mineração de Textos (*Text Mining*) podem ampliar a compreensão das tendências de negócios e auxiliar na identificação de riscos, contribuindo para a definição de ações que fidelizem os clientes.

Uma fonte valiosa para a descoberta de conhecimento a respeito de cliente se encontra nos dados em formato textual. No entanto, esse tipo de dado geralmente não é utilizado pelas organizações de negócio, devido à sua dificuldade em lidar com esse tipo de informação. A grande quantidade de dados textuais disponível nas bases de dados dificulta a sua organização, análise e extração. Porém, com o uso de técnicas automáticas de Mineração de Textos, esses processos se tornem viáveis [10, 31]. A descoberta de conhecimento em textos é análoga à descoberta de conhecimento em bases de dados [11], diferenciando-se, fundamentalmente na fase do pré-processamento, já que nos textos, os dados ainda se encontram em formato não-estruturado e, portanto, devem ser submetidos a um processo de estruturação [12].

Empresas de telefonia podem se beneficiar das técnicas de Mineração de Textos, já que essas organizações possuem campos textuais presentes em alguns bancos de dados das centrais de atendimento telefônico ao cliente. Nesses campos, são relatadas as ocorrências efetuadas pelos clientes, as quais podem conter informações relevantes sobre suas experiências e preferências. Desse modo, a organização pode analisar os requisitos dos clientes e utilizá-los como critérios para a reavaliação do seu desempenho no mercado. Essa compreensão possibilitará o desenvolvimento de modelos de desempenho estratégico e de análises de lacunas processuais de negócios, fundamentais para a tomada de decisões.

O presente trabalho teve como objetivo a aplicação de técnicas de Mineração de Textos para identificar os clientes que, potencialmente, poderiam desistir do plano de uma operadora de telecomunicações, e utilizar o conhecimento a respeito desses clientes para uma aplicação de gestão do conhecimento, com vista a uma potencial melhoria na gestão do relacionamento com os clientes. Para a aplicabilidade da pesquisa, a metodologia utilizada seguiu uma abordagem baseada na classificação de clientes em conectados ou desconectados, e também na utilização de um modelo de representação baseado em regras de associação que possibilitou a extração dos principais atributos que definem essas duas classes. Esses atributos extraídos foram então utilizados para avaliar o nível de satisfação dos clientes com relação aos serviços prestados pela empresa.

Este artigo está estruturado da seguinte maneira: após esta introdução, segue a Seção 2, que traz os trabalhos relacionados ao tema da pesquisa. A Seção 3 apresenta a proposta do trabalho, seguida pela Seção 4, que traz os experimentos e os resultados. Por fim, a Seção 5 mostra a conclusão.

2. TRABALHOS RELACIONADOS

Existe uma enorme quantidade de dados textuais em bancos de dados organizacionais que podem ser usados para coletar *feedback* explícito dos clientes. Essas informações possibilitam a extração de padrões e conhecimento a respeito desses clientes, úteis para o desenvolvimento de melhores estratégias de mercado. No entanto, a complexidade do relacionamento com o cliente torna a medição das suas percepções sobre experiências de serviços desafiadoras [13]. Para sanar esse problema, utiliza-se técnicas de Mineração de Textos [14], que auxiliam na identificação de padrões de clientes à partir de dados textuais.

Barcelos *et al.* [15] aplicaram *softwares* de categorização para a análise de dados não estruturados no contexto da gestão de clientes com relação a hotéis de um destino turístico. A aplicação foi realizada em comentários presentes em páginas de planejamento de viagens. Os dados foram coletados automaticamente, pré-processados, categorizados por temas e disponibilizados em visualizações interativas e dinâmicas para análises qualitativas e quantitativas dos comentários, condensando as informações disponíveis. Os resultados demonstraram que a mineração de textos e dados pode contribuir para se traçar estratégias de planejamento e gestão, além de auxiliar na identificação de problemas.

Dados estruturados e não estruturados também foram integrados para descobrir conhecimento, em Fatudimu *et al.* [17]. No processo de integração, o componente estruturado foi selecionado com base nas palavras-chave decorrentes do processo de pré-processamento de texto e regras foram geradas com base em um algoritmo baseado em regras de associação. O componente não estruturado da integração baseia-se na técnica de recuperação da informação que, por sua vez, se baseia, na semelhança do conteúdo de XML (*Extensible Markup Language*) do documento. Esta semelhança é baseada na combinação de relevância sintática e semântica dos termos, revelando uma significativa redução dos grandes conjuntos de itens e do tempo de execução. As experiências realizadas revelaram que as regras de associação extraídas contêm características importantes que constituem uma plataforma digna para tomar decisões eficazes no que diz respeito à gestão de relacionamento com o cliente.

Em Ordenes *et al.* [13], foi utilizada a Mineração de Textos baseada em modelagem linguística para a criação de um *framework*. Esses autores utilizam uma abordagem holística para analisar o *feedback* do cliente, incorporando elementos da experiência do cliente, metodologias e teorias de serviços. Para a mineração de texto baseada em linguística fizeram uso de recursos como bases de dados on-line de termos linguísticos (por exemplo, dicionário, dicionário de sinônimos). Para o processo de classificação automatizada de comentários em elogios e reclamações, o modelo incluiu padrões linguísticos baseados em subcategorias que diferenciavam os elogios das queixas. Os resultados empíricos mostraram que o modelo de Mineração de Textos para análise do *feedback* textual do cliente possibilita às empresas avaliar o impacto dos processos de serviços interativos sobre as experiências desses consumidores, auxiliando nos processos de co-criação bem-sucedida em um ambiente de serviço.

Takeuchi e Yamaguchi [19] aplicaram técnicas de Mineração de Textos em transcrições de registros textuais de conversas telefônicas pertencentes a uma central de atendimento ao cliente de um escritório de reservas de veículos alugados. Para isso, identificaram expressões importantes usadas pelos clientes, às quais atribuíram categorias semânticas usando um dicionário com base no conhecimento de um especialista do domínio. Essas expressões, ou palavras-chave, estão relacionadas a categorias, como "produto" e "problema". O método foi utilizado para analisar quase mil conversas, buscando novas ideias para melhorar a produtividade dos funcionários, o que resultou em um aumento real das receitas.

Cailliau e Cavet [20], a fim de assegurar e melhorar a qualidade do serviço nas centrais de atendimento telefônico, utilizaram um método para selecionar e classificar conversas críticas através da Mineração de Textos. Utilizaram abordagem linguística, baseada em análise de sentimentos, para detectar marcadores de sentimento em transcrições de fala automática francesa. O peso e a orientação dos marcadores foram usados para calcular a orientação semântica da fala, segundo critérios de classificação dos termos em classes com orientação positiva, negativa ou neutra (Aceitação - Recusa, Acordo - Desacordo, Favorável - Apreciação desfavorável, Opinião, Surpresa). O curso de uma conversa pôde então ser representado graficamente com curvas positivas e negativas. Os autores avaliaram um corpus manualmente anotado e utilizaram heurísticas para a seleção automática de conversas problemáticas. Tais heurísticas provaram ser muito úteis e complementares para a recuperação de conversas com segmentos de raiva e tensão.

Todos esses trabalhos revelam a importância que os processos de Mineração de Textos assumem no contexto da extração e representação do conhecimento para sua aplicação na gestão de relacionamentos com o cliente. No entanto, na maioria das vezes, os métodos utilizados ainda apresentam forte dependência de questionários, dicionários de termos e/ou de conhecimento de um especialista do domínio para a elaboração de uma lista pré-definida de atributos de interesse para o processo de Mineração de Textos, o que prejudica o seu desempenho em termos de tempo e de custo.

O modelo de representação utilizado nesse trabalho utiliza como atributos palavras relacionadas que se repetem na própria coleção de textos. A análise é executada em espaços limitados, ao longo de cada um dos documentos, ao invés de palavras isoladas, além de não necessitar do auxílio de uma lista pré-definida de atributos do domínio. Dessa maneira mantém, na medida do possível, a carga semântica (significado) original dos textos analisados, a fim de aplicar o conhecimento extraído em melhorias na gestão organizacional e no relacionamento com os clientes em uma empresa de telecomunicações.

3. TRABALHO PROPOSTO

Apresenta-se, nesse trabalho, uma abordagem da Mineração de Textos baseada na classificação de clientes e na extração de regras de associação que revelem suas percepções e experiências com relação aos serviços prestados por uma empresa de telecomunicações. Dada uma base de dados, contendo ocorrências diversas, registradas por operadores do serviço de atendimento ao cliente por telefone, objetivamos processar os dados textuais e extrair padrões que permitam aos gestores antever a ocorrência de cancelamentos dos planos, além de auxiliar na formulação de ações estratégicas que contribuam para a fidelização e a permanência do cliente no plano.

Para isso, foram adotadas duas abordagens de Mineração de Textos: a classificação de clientes em duas classes (conectados/desconectados) e a extração de regras de associação para a extração de atributos pertencentes a cada uma das classes. Para a classificação das ocorrências entre aquelas que tendem a resultar em cancelamento do plano da operadora e aquelas que indicam possibilidade de permanência no plano por parte do cliente, foi criada uma representação de frequência dos termos, organizada em uma *Bag-Of-Words* convencional [28]. Essa representação foi posteriormente submetida a algumas técnicas de classificação.

Para a análise de percepções e preferências dos clientes, a abordagem utilizada foi baseada na extração de termos compostos, representados em uma *Bag-Of-Related-Words* [23], gerada pela aplicação da ferramenta *Features gEnerator based on AssociaTion RulEs* (FEATuRE) [21]. O objetivo desse modelo de representação é utilizar como atributos palavras relacionadas que se repetem, em espaços limitados, ao longo de um documento ao invés de palavras isoladas. Dessa maneira, busca-se um modelo de representação que mantenha, na medida do possível, a carga semântica (significado) dos textos analisados a fim de aplicar o conhecimento extraído em melhorias na gestão organizacional e no relacionamento com os clientes da empresa.

3.1 Coleção de Dados Não-Estruturados

O conjunto de dados não estruturados usado para o experimento foi cedido por uma empresa de telecomunicações e contém todas as ocorrências registradas em formato textual, dentro de um período de 12 meses, formando um conjunto específico de 12.990 ocorrências de 152 clientes, classificados em conectados e desconectados. Os clientes desconectados cancelaram o serviço dentro do período de 1 ano e representam a metade do conjunto de documentos.

O conteúdo de uma ocorrência segue a estrutura ilustrada na Figura 1, na qual informações de local, data, horário e cliente são detalhadas. No cabeçalho, constam informações geradas automaticamente pelo sistema de atendimento. A ocorrência pode restringir-se apenas a essas informações nos casos em que não houver atendimento pessoal. Nesses casos a transação é gerada somente pelo sistema automático. Do contrário, ao fim da seção, segue a informação digitada pelo atendente, em formato textual. Trata-se do material de interesse para este trabalho.

DADOS CONFIRMADOS: SIM (X) NÃO ( )
PROTOCOLO:SIM (X) NÃO ( )
CARGO\SETOR:
TELEFONE:
EMAIL

O CLIENTE LEONARDO COMELLI, CPF 99.999.999-18 CONFIRMOU OS DADOS E RELATA ESTAR SEM SINAL DE INTERNET E TELEFONE JÁ FEITO TODOS OS TESTE SINAL NÃO NORMALIZADO.

REGIÃO DO CLIENTE ESTAR COM OUTAGE PRAZO
DE RETORNO 15/09/2016 HORAS: 16:00
DITO POR CAETANO
DILDRE \ JOÃO \ XPTO EMPRESAS \ YTZ RECIFE

Figura 1. Exemplo hipotético de ocorrência.

3.2 Pré-processamento

O Pré-processamento foi aplicado para melhorar a qualidade dos dados, reduzindo a quantidade de ruídos, valores discrepantes (*outliers*), inconsistências ou qualquer outro fator que possa comprometer a eficiência das etapas de mineração subsequentes. Uma base de dados construída a partir de transcrições realizadas por atendentes com base nos relatos de clientes de uma empresa de telecomunicações é altamente suscetível a ruídos. Isso se deve às dificuldades de comunicação decorrentes da localização dos atendentes e dos clientes, que podem estar espalhados por diferentes regiões do país, além de possuírem níveis distintos de conhecimento, escolaridade e cultura. Aplicando-se algumas

técnicas de limpeza dos dados antes do processo de mineração é possível melhorar, substancialmente, a qualidade dos padrões minerados e o tempo necessário para a mineração [22].

Desse modo, o fluxo de pré-processamento foi dividido em três passos, conforme ilustrado na Figura 2. O primeiro passo diz respeito à Preparação dos Textos e consistiu na conversão de todos os textos para caixa baixa, remoção dos dados confidenciais dos atendentes e clientes (nomes, número de documentos, telefones, etc.), remoção de padrões específicos do domínio (úteis apenas para o processo de atendimento e não relevante para a mineração), limpeza (incluindo a remoção dos caracteres especiais, dígitos, pontuação, acentos e quebra de linha), e remoção de algumas *stopwords* (artigos, preposições, conjunções).

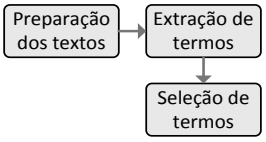

Figura 2. Fluxo do pré-processamento.

O segundo passo foi a Extração de Termos candidatos a atributos. Visando a normalização dos textos que compõem a coleção de documentos e a redução da matriz de atributos, foi realizado o processo de radicalização (*stemming*) de cada um dos termos. Esse processo consistiu em manter apenas os radicais das palavras (denotação mínima e não ambígua do termo). Após obter o radical de cada termo, uma representação numérica da coleção de documentos, do tipo *Bag-Of-Words* [28], foi gerada. Esse tipo de representação se insere no Modelo Espaço Vetorial, no qual cada documento é representado por um vetor e cada posição desse vetor corresponde a dimensões (atributos). Essa representação foi então aplicada à primeira etapa do trabalho (Subseção 3.3), isto é, na classificação de ocorrências que podem ocasionar cancelamentos.

Para concretizar a primeira etapa deste trabalho (classificação), foi incluído um último passo, que diz respeito à Seleção de Termos. Para diminuir a dimensão da matriz da atributos, foram avaliadas as frequências do termos no documento e na coleção de documentos. Os termos com maior relevância compuseram a nova matriz de atributos. Tanto a matriz completa como a matriz reduzida foram submetidas a diferentes técnicas de Aprendizado de Máquina, as quais constituíram o processo de Mineração de Dados propriamente dito.

Para a segunda etapa do trabalho (gestão de clientes, Subseção 3.4), a extração de termos seguiu um processo diferente, objetivando uma representação que mantivesse a posição relativa das palavras nos textos e reportasse também os termos compostos a fim de preservar a compreensibilidade dos termos extraídos sem prejudicar a representatividade do domínio. A abordagem *Bag-Of-Words* não considera as informações sintáticas e semânticas que estruturam um texto, assumindo que a frequência das palavras é uma variável independente, além de desprezar o fato de que muitos termos são compostos por duas ou mais palavras, úteis para o processo de Mineração de Textos e interpretação dos resultados. Outro problema encontrado nesse tipo de representação diz respeito à alta dimensionalidade e esparsidade da matriz apresentada mesmo após a etapa de pré-processamento.

Para evitar estes problemas, o conjunto de documentos foi processado pela ferramenta FEATuRE [21], capaz de gerar atributos compostos por mais de uma palavra e representá-los em uma *Bag-Of-Related-Words*. O objetivo desse modelo de representação é utilizar como atributos palavras relacionadas que se repetem, em espaços limitados, ao longo de um documento [21, 23]. As palavras relacionadas são obtidas por meio de regras de associação [24], que extraem relações entre itens em uma base de dados na qual A => B, significando que quando um termo A ocorre, um termo B também tende a ocorrer. Com as regras obtidas no processamento, foram extraídos os termos com maior índice de frequência nos documentos de cada um dos conjuntos de documentos (clientes conectados/clientes desconectados), os quais, por sua vez, foram avaliados em conjunto com os termos que se ligam a eles. Posteriormente, esses termos foram comparados aos atributos de qualidade, sob a perspectiva do cliente, presentes na literatura [25]. Com isso, foi gerada uma Matriz de Importância e Desempenho, indicando a posição de cada atributo/critério de desempenho em relação à sua importância dada pelo cliente e ao desempenho atual relativo à empresa prestadora de serviço.

3.3 Classificação

A fim de classificar as ocorrências em "cliente conectado" ou "cliente desconectado" e gerar um modelo capaz de predizer cancelamentos, diferentes técnicas de Aprendizado de Máquina foram aplicadas, em contexto supervisionado e semi-supervisionado. No contexto supervisionado, foram avaliadas técnicas de diferentes naturezas, sendo considerados os algoritmos J48 para Árvores de Decisão, RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) para extração de Regras de Decisão, Rocchio para modelos espaço vetorial, e KNN (*k-Nearest Neighbours*), para aprendizado baseado em instâncias. No contexto semi-supervisionado, as principais abordagens foram os algoritmos TSVM (*Transductive Support Vector Machine*) e TCHN (*Transductive Classification based on Heterogeneous Network*).

Os testes foram realizados por meio de 10-*fold cross-validation*, sendo que em cada uma das dez iterações do procedimento uma porção foi separada para testar a indução supervisionada ou semi-supervisionada e as nove porções restantes foram utilizadas para o treinamento. Esse procedimento visa selecionar partições diferentes para os treinamentos e os testes, para assim se chegar a um resultado mais confiável e evitar a seleção de apenas uma parte do corpus, o que poderia influenciar na execução dos algoritmos.

3.4 Gestão de Clientes

Conhecer o cliente constitui a base da Gestão de Relacionamento com Clientes [29]. Essa abordagem busca a fidelização em decorrência do bom relacionamento. Tal relacionamento pode auxiliar a organização no que diz respeito à preferências e possibilidades de aquisição e manutenção de novos clientes, direcionando assim a sua estratégia para uma maior lucratividade. Portanto, a organização de negócios deve analisar os requisitos dos clientes e utilizá-los como critérios para a avaliação de desempenho de seus processos internos. Todavia, esses modelos são influenciados e orientados pelas informações dos clientes, que primeiramente devem ser extraídas e analisadas. Desse modo, a organização poderá avaliar a satisfação dos clientes e a diferença entre seu desempenho real e o desempenho ideal [26]. Essa análise, além de gerar conhecimento, auxiliará a organização na

promoção das devidas melhorias e garantirá a entrega de valor a seus clientes.

Para auxiliar nessa tarefa, foi criada a Matriz de Importância e Desempenho [27], também conhecida como Matriz de Qualidade e Desempenho, que indica a posição de cada critério de desempenho em relação à sua importância e ao seu desempenho atual. Tudo isso direcionado, primeiramente, à definição da importância de cada um dos critérios. Cada atributo de um serviço pode ser julgado pela importância dada pelo consumidor e pelo desempenho da empresa. Nesse sentido, Sias (2005), identificou e classificou, por meio de uma pesquisa quantitativa, nove atributos de qualidade nos serviços de uma empresa brasileira de telecomunicações, percebidos pelos clientes e classificados quanto à sua importância, em ordem decrescente, como pode ser visto na Tabela 1 [25].

É através da classificação dada pelos clientes que os atributos de qualidade são ordenados em ordem do "mais importante" ao "menos importante", ou seja, quanto mais próximo ao 1, mais importante é esse atributo para o cliente. Esses atributos estão vinculados a algumas variáveis, classificadas pelos consumidores das empresas de telecomunicações como as mais importantes para a sua satisfação e permanência no plano.

Tabela 1. Classificação dos requisitos de qualidade segundo sua importância para o cliente

Valor	Atributo	Variáveis
1	Preço	Mensalidade cobrada
2	Atendimento	Atendimento do consultor
3	Segurança	A garantia do serviço
4	Confiabilidade	O desempenho da velocidade contratada
5	Qualidade	A qualidade do serviço
6	Instalação	Tempo de instalação
7	Reparo	Tempo de reparo
8	Funcionários	Funcionário de instalação e reparo
9	Conveniência	Horários de instalação e manutenção

4. EXPERIMENTOS E RESULTADOS

O desenvolvimento desse trabalho se deu mediante dois experimentos: a classificação de clientes em "conectados" e "desconectados" e a análise dos requisitos de qualidade extraídos das ocorrências de cada uma das classes de clientes. Esses dois experimentos são descritos a seguir.

4.1 Experimento: Classificação de Clientes

A partir da aplicação da etapa de Pré-processamento sobre o conjunto de textos, foi gerada uma *Bag-Of-Words*, com um total de 10933 termos. Em seguida, foi aplicada a métrica TD-IDF (*Term Frequency – Inverse Document Frequency*) para determinar a relevância de cada termo dentro da coleção de documentos. A aplicação das técnicas de Aprendizado de Máquina, com diferentes parâmetros, possibilitou a classificação dos dados com acurácia. A Tabela 2 mostra os principais resultados obtidos com a aplicação das técnicas de classificação.

Os melhores resultados foram obtidos com aplicação no contexto supervisionado, particularmente em Árvores de Decisão resultantes do algoritmo J48 e nas Regras de Decisão provenientes do algoritmo RIPPER. Com estas técnicas, foram obtidos 95,69% e 97,02% de acurácia, respectivamente.

Tabela 2 - Principais resultados obtidos com aplicação das técnicas de classificação

Algoritmo	Acurácia média	Desvio-padrão
J48	95,69%	4,83
RIPPER	97,02%	4,92
Rocchio	80,63%	13,04
KNN	81,85%	10,09
TSVM	88,52%	5,80
TCHN	88,10%	8,09

O algoritmo de aprendizado baseado em instâncias KNN levou a uma acurácia de até 81,85% para k igual a três utilizando a medida de proximidade cosseno. Algoritmos de aprendizado semi-supervisionado também levaram a resultados promissores. Com aplicação do TSVM, chegou-se a obter 88,52% de acurácia, tomando 50 exemplos na etapa supervisionada, e até 87,13% ao se tomarem 40 exemplos nesta etapa. Já, com o algoritmo TCHN, chegou-se a obter 88,10% de acurácia.

Esses resultados são compatíveis com o estado da arte na área de predição de cancelamentos de contratos em empresas de telecomunicações, em alguns casos até superando esses resultados [30].

4.2 Requisitos para Gestão de Clientes

Para que os algoritmos de Mineração de Textos possam ser manipulados, os dados brutos devem, primeiramente, ser representados de modo apropriado. Para isso, nesse trabalho foi selecionada a ferramenta FEATuRE [21], capaz de gerar atributos compostos por mais de uma palavra e representá-los em *Bag-Of-Related-Words*.

Na fase de Pré-processamento, optou-se por realizar a remoção de *stopwords* e a radicalização de palavras, que possibilitou a substituição das palavras radicalizadas pela palavra mais frequente que originou o radical. Esses procedimentos possibilitaram a diminuição da dimensionalidade e da esparsidade da matriz. Posteriormente, foram extraídas as regras de associação, mapeando cada documento em um conjunto de transações. Para isso utilizou-se a opção de janelas deslizantes como transações, pois essas janelas correspondem a trechos do documento textual e extraem conjuntos de palavras que estão relacionadas em um contexto específico desse documento. Nesse tipo de mapeamento, a primeira transação contém apenas a primeira palavra do documento, a segunda contém as duas primeiras palavras, e assim por diante, até que a janela contenha o número de palavras igual ao tamanho definido para a janela, que no presente trabalho foi definido em 5 com salto em 1, já que essas medidas foram capazes de capturar o sentido de interesse das palavras presentes nos textos.

Para a extração das regras de associações das transações mapeadas, foi definido o valor do limiar de suporte mínimo automático. A fórmula para o cálculo do suporte mínimo leva em consideração a frequência média das palavras nas transações. Esse valor isenta o usuário de conhecer as características do documento ou da coleção de documentos, além de evitar a definição de um valor de suporte mínimo muito baixo, que poderia gerar uma quantidade muito grande de regras [21].

Para a representação *Bag-Of-Related-Words,* utilizou-se os itens das regras de associação para compor os atributos, definindo o tipo *itemsets* (conjunto de itens frequentes), com limiar de medida objetiva automático, visando a diminuição da dimensionalidade da representação. O resultado de interesse obtido nesse processo foi uma lista dos principais termos simples e compostos extraídos dos dois conjuntos de documentos, relativos aos clientes conectados e aos clientes desconectados, como mostra a Tabela 3.

Tabela 3. Atributos extraídos no processo de mineração.

Clientes conectados	Clientes desconectados
aberto-contrato-dívida	aberto-contrato-dívida
acordo-pagamento	acordo-pagamento
aguardar-chamado-regularização	aguardar-chamado-regularização
atraso-pagamento	atraso-pagamento
cliente-informações-solicita	cliente-informações-solicita
cliente-informações-protocolo-solicita	cliente-informações-protocolo-solicita
cliente-protocolo-solicita	cliente-protocolo-solicita
dívida-possui	dívida-possui
recuperação-valor	recuperação-valor
pagamento-promessa	pagamento-promessa
cobrança-duplicidade-motivo	arquivo-cobrança
cobrança-motivo-valor	erro-motivo-valor
dívida-possui-reversão	erro-valor
duplicidade-motivo-valor	dívida-negociada
corrigida-inconsistência-notas	boleto-cobilling-processo
corrigida-inconsistência-telefone	cobrança-processo
contrato-executar-possível-reversão	cobrança-suspensão
contrato-possui-reversão	ocorrência-ongoing
informações-protocolo-solicita	desk-service-solicitando
dívida-reversão	devido-equipamento-usuário
inconsistência-nota	pagamento-previsão
motivo-valor	processo-retirado

É possível notar que os atributos de destaque presentes nas duas classes de clientes (conectados/desconectados) da Tabela 3 são compostos pelos termos "dívida", "pagamento", "valor", "cobrança". Esses termos dizem respeito ao critério "***Preço***", corroborando com os resultados apresentados na literatura [25]. Isso indica que a empresa deve se atentar a esse requisito, já que na escala de importância ele recebe o valor de número 1, ou seja, é o mais importante na lista de qualidades do cliente.

Outro problema encontrado em ambas as classes de clientes diz respeito a "erros no valor da fatura" por motivos de "inconsistência" e "erros de cobrança". Esses itens também estão relacionados ao critério ***"Preço"***, o que também pode gerar desconforto e apreensão nos clientes com relação aos critérios**,** ***"Atendimento", "Segurança" e "Confiabilidade"***, com importância de valor número 2, 3 e 4, respectivamente.

No grupo dos clientes desconectados, o atributo "equipamento devido ao usuário", ligado aos critérios **"Segurança"**, **"Qualidade"** e **"Reparo"** também ganha destaque, pois só aparece nesse grupo, indicando que pode exercer um peso importante na decisão do cliente de se desligar do plano e, em decorrência disso, deve receber adequada atenção dos gestores da organização. Esses critérios apresentam importância de valor de número 3, 5 e 7, respectivamente.

Um critério que também se destacou entre os atributos de ambas as classes de clientes e que não apareceu na literatura como critério de avaliação, diz respeito ao atributo "***Informação***". Isso significa que o processo foi capaz de extrair um requisito de qualidade para os clientes que não havia sido detectado anteriormente. Essa informação não apareceu nas pesquisas aplicadas na detecção dos principais atributos sob a ótica do cliente [25], mas foi detectada nesse experimento.

Com essas informações foi possível construir a Matriz de Importância e Desempenho, como mostra a Figura 3. Essa matriz indica aquilo que deve ser priorizado no ataque ao aprimoramento do desempenho dos critérios competitivos, devendo-se estabelecer planos de ação para atingir o aprimoramento desejado [26]. A matriz é dividida em quatro zonas distintas: i) zona adequada (onde o desempenho atual é considerado satisfatório), ii) zona de melhoria (onde o desempenho se encontra abaixo do aceitável e precisa ser melhorado, iii) zona de urgência (onde o desempenho de um critério muito importante para o cliente se encontra muito abaixo do aceitável, exigindo ações de melhorias imediatas) e, iv) zona de excesso (onde o desempenho é alto demais em relação à importância que o cliente dá esse critério).

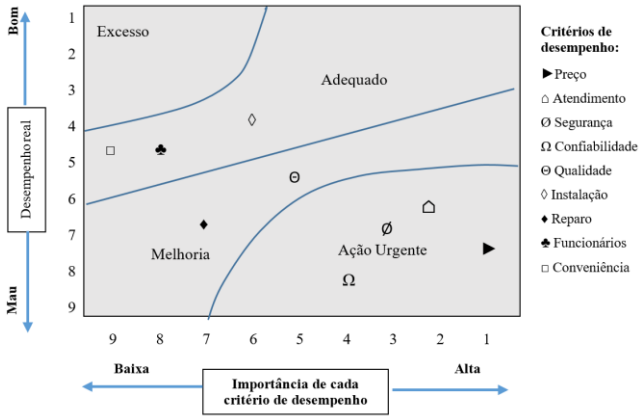

Figura 3. Matriz de Importância e Desempenho

A matriz é também delimitada por uma linha diagonal que indica o "limite de aceitabilidade" entre o desempenho atual "aceitável" e o "inaceitável". Todavia, vale ressaltar que, nem sempre o critério que se encontra abaixo desse limite exige esforços e prioridade de melhoria. Isso ocorre devido ao caráter relativo à importância do critério, que varia conforme o mercado ou segmento de mercado focado pela organização [26]. Por isso, a organização deve extrair informações do cliente e analisar quais são os critérios válidos para a análise de seus processos e estipular a importância relativa de cada um deles.

O resultado da análise matricial pode ser resumido na Tabela 4, composto por três colunas. A primeira coluna diz respeito ao valor da "Importância" de cada critério segundo a classificação dos clientes, a segunda coluna traz a especificação dos próprios "Critérios de desempenho" e seus respectivos símbolos (presentes na matriz da Figura 3 e, por fim, a terceira coluna traz os valores do "Desempenho real" da organização com relação aos critérios estabelecidos pelos clientes.

Com esse resultado, a organização poderá investir em ações estratégicas específicas para a resolução dos problemas aqui

apontamos, de modo a buscar soluções eficientes para resolver tais problemas e manter os clientes fiéis ao plano. Os resultados identificados nessa etapa, possibilitaram verificar que os critérios de número 1 a 4 se encontram na zona de "Ação Urgente", exigindo ações estratégicas imediatas para reverter essa situação, considerada crítica para a gestão do relacionamento com os clientes. Os critérios de número 5 e 7 também exigem melhorias a médio prazo, pois tendem a decair no decorrer do tempo, já que se encontram na "zona de melhoria" a qual aponta que o desempenho se encontra abaixo do aceitável e precisa ser melhorado. Com relação aos critérios de importância número 6 e 8, não há necessidade de mudanças, pois se encontram na zona adequada, com desempenho atual considerado satisfatório.

Tabela 4. Resultado da Matriz de Qualidade e Desempenho

Valor	Critérios de desempenho	Desempenho real
1	► Preço	8
2	⌂ Atendimento	7
3	Ø Segurança	7
4	Ω Confiabilidade	8
5	Θ Qualidade	6
6	◊ Instalação	4
7	♦ Reparo	7
8	♣ Funcionários	5
9	□ Conveniência	5

Essa análise foi efetuada com base nos atributos que mais se destacaram no processo de extração de regras e na representação *Bag-Of-Related-Words*. Os resultados obtidos foram satisfatórios, já que corroboraram com o trabalho de Sias [25]. Ambos os trabalhos convergiram no que diz respeito à extração dos principais critérios de qualidade na área de negócios em telecomunicações sob a ótica do cliente. No entanto, nesse trabalho surgiu o critério "Informação", que não fora identificado nos questionários aplicados por Sias [25] e que demonstra ser importante para os clientes, pois aparece entre os principais atributos extraídos pelas regras de associação.

A principal vantagem da Mineração de Textos é que ela é realizada automaticamente, podendo utilizar conhecimento disponível em bases de dados já existentes, comprovando o potencial papel desses bancos de dados como fonte de informação e conhecimento. A técnica automatizada elimina a necessidade de maiores interações com o usuário e a aplicação de questionários. Isso implica em satisfação do cliente, pois não é necessário abordá-lo diretamente para responder perguntas referentes à sua satisfação.

A abordagem proposta neste trabalho é independente de domínio, e por esse motivo não necessita da utilização de dicionários e de listas de termos pré-confeccionadas, proporcionando uma economia de tempo e de custos, além de possibilitar a descoberta de novos atributos importantes, até então desconsiderados que, por sua vez, auxiliam na geração de novos conhecimentos.

Como resultado, as técnicas aqui propostas extraíram informações relevantes, capazes de gerar conhecimento a respeito dos clientes e do desempenho da organização. Tal conhecimento pode contribuir para a tomada de decisões estratégicas e eficazes no que diz respeito à gestão de relacionamento com o cliente. Portanto, pode-se afirmar que as técnicas aqui utilizadas contribuem para a identificação de atributos relacionados aos requisitos de clientes, destacando a importância da abordagem da Mineração de Textos para a extração de padrões e de conhecimento.

5. CONCLUSÕES

Este trabalho abordou ocorrências de clientes registradas por funcionários do serviço de atendimento por telefone de uma empresa de telecomunicações, aplicando-se técnicas de Mineração de Textos e Aprendizado de Máquina. Duas abordagens distintas foram utilizadas para a extração de dados: na primeira foi analisada, por meio do processo de classificação, a capacidade preditiva desses dados em relação à probabilidade de cancelamentos do plano. Na segunda abordagem foi realizada uma análise do ponto de vista gerencial, tomando por base a extração de termos e as relações entre eles, baseando-se em regras de associação para mostrar que as informações extraídas podem ser utilizadas no auxílio à tomada de decisões estratégicas para melhorar a gestão de clientes.

Do ponto de vista da classificação das ocorrências, os melhores resultados foram obtidos com técnicas baseadas em regras de Indução ou Árvores de Decisão, com os algoritmos RIPPER e J48 respectivamente. Do ponto de vista gerencial, a extração de regras de associação gerou conjuntos de termos/atributos úteis para aplicações de ferramentas de gestão do conhecimento organizacional. Com os resultados obtidos nas duas abordagens foi possível realizar uma análise de padrões que podem levar um cliente a permanecer ou a se desvincular do plano da operadora. Esses resultados podem ser utilizados pelos gestores, com vista à adoção de melhores estratégias de negócio à fim de aumentar a satisfação dos clientes e garantir sua fidelidade em um mercado competitivo.

A limitação da aplicação aqui proposta se encontra na ausência de uma abordagem de cunho semântico para a extração de padrões, capazes de justificar as associações (ou relacionamentos) entre os termos que compõem os atributos compostos. Portanto, trabalhos futuros podem se focar no desenvolvimento de métodos que introduzam análises semânticas para justificar os relacionamentos e refinar os dados extraídos, agregando confiabilidade aos resultados obtidos no processo de Mineração de Textos.

6. AGRADECIMENTOS

Este trabalho recebeu financiamento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2017/02377-5.

7. REFERÊNCIAS

[1] Prahalad, C. K., & Krishnan, M. S. (2008). *The new age of innovation: Driving cocreated value through global networks*. McGraw Hill Professional.

[2] Angelo, C. F. D., & Giangrande, V. (1999). Marketing de relacionamento no varejo. *São Paulo: Atlas*, 10.

[3] Swift, R. (2001). Customer Relationship Management-O Revolucionário Marketing de Relacionamentos com Clientes.

[4] Kloter, P., & Armstrong, G. (2003). Fundamentos de marketing. *México: 6ta Edición Prentice Hall*.

[5] Oliveira, C. P., Leão, M. C. S., & Costa, R. A. T. (2016). Gestão do Relacionamento com os clientes: Um Estudo na

Agência Beira Rio do Banco do Brasil. *Revista de Administração Geral*, *1*(2), 21-40.

[6] Turban, E., Sharda, R., Aronson, J. E., & King, D. (2009). *Business Intelligence: um enfoque gerencial para a inteligência do negócio*. Bookman Editora.

[7] Turban, E., Leidner, D., Mclean, E., & Wetherbe, J. (2010). *Tecnologia da Informação para Gestão-: Transformando os Negócios na Economia Digital*. Bookman.

[8] Beránková, M. H., & Houska, M. (2011). Data, information and knowledge in agricultural decision-making. *AGRIS on-line Papers in Economics and Informatics*, *3*(2), 74.

[9] Piatetsky-Shapiro, G. (1996). *Advances in knowledge discovery and data mining* (Vol. 21). U. M. Fayyad, P. Smyth, & R. Uthurusamy (Eds.). Menlo Park: AAAI press.

[10] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[11] Hearst, M. A. (1999, June). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10). Association for Computational Linguistics.

[12] Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.

[13] Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing Customer Experience Feedback Using Text Mining A Linguistics-Based Approach. *Journal of Service Research*, 1094670514524625.

[14] Zhong, N., Li, Y., & Wu, S. T. (2012). Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering*, *24*(1), 30-44.

[15] de Barcelos, Y. T., Gosling, M., de Freitas Coelho, M., & Resende, M. P. (2014). Ferramenta de visualização de dados e processamento de texto: análise de reviews de viajantes no Tripadvisor. *REAVI-Revista Eletrônica do Alto Vale do Itajaí*, *3*(4), 25-39.

[17] Fatudimu, I. T., Uwadia, C. O., & Ayo, C. K. (2012). Improving Customer Relationship Management through Integrated Mining of Heterogeneous Data. *International Journal of Computer Theory and Engineering*, *4*(4), 518.

[18] Jain, A. K., Murty M. N., & Flynn, P. J. (1999). *Data clustering: a review*, ACM Computing Surveys, vol. 31, no. 3, pp. 264–323.

[19] Takeuchi, H., & Yamaguchi, T. (2014). Text mining of business-oriented conversations at a call center. In *Data mining for service* (pp. 111-129). Springer Berlin Heidelberg.

[20] Cailliau, F., & Cavet, A. (2013, March). Mining automatic speech transcripts for the retrieval of problematic calls. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 83-95). Springer Berlin Heidelberg.

[21] Rossi, R. G., & Rezende, S. O. (2011a). Generating features from textual documents through association rules. *Anais do Encontro Nacional de Inteligência Artificial. SBC*. São Carlos – SP, Brasil.

[22] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[23] Rossi, R. G., & Rezende, S. O. (2011b). Building a topic hierarchy using the bag-of-related-words representation. In *Proceedings of the 11th ACM symposium on Document engineering* (pp. 195-204). ACM.

[24] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

[25] Sias, C. C. (2005). O desempenho dos atributos de qualidade em serviços de conectividade de redes: o caso de uma operadora de telecomunicações. Mestrado - [s.l.] Universidade Federal do Rio Grande do Sul.

[26] Carvalho, F. C. A. (2012). *Gestão do Conhecimento*. São Paulo. Person.

[27] Slack, N., Chamners, S., Harland, C., Harrison, A. & Johnston, R. (1999). Administração da Produção. São Paulo. Atlas.

[28] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). ACM.

[29] Silva, F. D., & ZAMBON, M. S. (2006). Gestão do relacionamento com o cliente. São Paulo: Thomson, 191.

[30] Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. *Computers & Electrical Engineering*, 38(6), 1808-1819.

[31] Lima, F., Oliveira, H., & Salvador, L. (2015). Um Método Não Supervisionado para o Povoamento de Ontologias a partir de Fontes Textuais na Web.