UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

COMBINAÇÃO DE DADOS ESTRUTURADOS E NÃO-ESTRUTURADOS PARA A MELHORIA DA PREDIÇÃO EM PROCESSOS DE NEGÓCIO

Matheus Augusto de Oliveira

Orientadores iniciais
Flavia Maria Santoro
Kate Cerqueira Revoredo

Orientador e revisor final
Pedro Nuno de Souza Moura

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2019

Catalogação informatizada pelo autor

O48 Oliveira, Matheus Augusto de
COMBINAÇÃO DE DADOS ESTRUTURADOS E NÃO-ESTRUTURADOS PARA A MELHORIA DA PREDIÇÃO EM PROCESSOS DE NEGÓCIO / Matheus Augusto de Oliveira. -- Rio de Janeiro, 2019.
63

Orientador: Pedro Nuno de Souza Moura.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2019.

1. Processos de negócio. 2. Mineração de processos. 3. Monitoramento preditivo. I. Moura, Pedro Nuno de Souza, orient. II. Título.

COMBINAÇÃO DE DADOS ESTRUTURADOS E NÃO-ESTRUTURADOS PARA A MELHORIA DA PREDIÇÃO EM PROCESSOS DE NEGÓCIO

Matheus Augusto de Oliveira

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado por:

_____

Pedro Nuno de Souza Moura

_____

Adriana Cesário de Faria Alvim

_____

María del Rosario Girardi Gutiérrez

RIO DE JANEIRO, RJ – BRASIL.

NOVEMBRO DE 2018

Agradecimentos

Agradeço a minha mãe, meu pai, meu irmão e irmã por todo esforço feito para que eu pudesse viver tudo o que vivi e os membros da minha família que, de diferentes maneiras, me deram suporte necessário durante esses anos de graduação.

Quero agradecer às professoras Kate e Flávia pela orientação inicial neste trabalho e ao professor Pedro pela orientação na etapa final. Obrigado por todas as dicas, opiniões, correções e principalmente pela paciência ao longo do desenvolvimento do projeto.

Quero agradecer também aos amigos que fiz ao longo dos anos de faculdade, pelos ótimos momentos vividos e experiências compartilhadas.

RESUMO

Os modelos de processos de negócio representam, de maneira rápida e visual, como as atividades desses processos estão organizadas, seus respectivos responsáveis e os impactos de cada uma delas. O objeto desses modelos é mostrar de forma eficiente e mais próxima possível da realidade, quais são os caminhos e os resultados existentes ao final de cada execução. Existem diferentes notações para modelização de processos e cada uma delas apresenta características específicas, que são traduzidas nos elementos utilizados para representar os diferentes atributos do processo. Dentro de uma organização, modelos bem definidos ajudam a organizar os fluxos existentes e possibilitam uma melhor alocação de recursos, uma vez que esses modelos conseguem traduzir o que acontece no dia a dia dessa organização e quem são os responsáveis pela execução a cada etapa.

A boa gestão dos processos de negócio abre caminho para outros temas, como por exemplo, a coleta das informações geradas a cada execução de um determinado processo e a aplicação de algoritmos que fazem o uso dessas mesmas informações para prever os resultados esperados em determinada instância do processo. Muitos desses algoritmos fazem uso de técnicas de aprendizado para encontrar padrões nos resultados do registro e aplicá-los no momento da execução de um novo caso para indicar seu resultado mais provável. Realizando um monitoramento em tempo real do processo e dessas previsões, ações podem ser tomadas com intuito de alterar esse resultado, caso o previsto não seja o esperado.

Através do uso desses padrões aprendidos, o monitoramento preditivo em processos de negócio visa a melhorar o grau de confiança dos resultados através de diferentes métodos. Esse monitoramento é feito pela combinação, aplicação e estudo das variáveis existentes no modelo do processo. O objetivo deste projeto de final de curso foi utilizar como base um estudo sobre como o uso de texto livre combinado aos outros atributos de modelos de processo melhora os resultados da previsão e aplicar configurações similares às usadas durante esse estudo em um outro contexto para reforçar a hipótese de que essa combinação apresenta melhor desempenho do que o uso de somente um dos tipos de informação. Os resultados mostram que, de modo geral, essa combinação apresenta melhor desempenho, mas que os algoritmos se comportam de maneira diferente de acordo com o contexto.

Palavras-chave: Processos de Negócio, Mineração de processos, Monitoramento preditivo.

ABSTRACT

Business process models represent how the processes activities are organized, the responsible and the impacts for each of them. The object of these models is to show as efficiently and as closely as possible to reality, what are the paths and the results at the end of each execution. There are different notations for process modeling and each of them has specific characteristics, which are translated into the elements used to represent the different attributes of the process. Within an organization, well-defined models help to organize the existing flows and enable a better allocation of resources, since these models can translate what happens in the organization daily tasks and who is responsible for the execution at each stage.

The good management of business processes opens the way to other topics, such as the collection of information generated at each execution of a given process and the application of algorithms that use the same information to predict the expected results in a given process case. Many of these algorithms make use of learning techniques to find patterns in the registry results and apply them during the execution of a new case to indicate its most likely outcome. By performing a real-time monitoring of the process and of these forecasts, actions can be taken with a view to changing this result, in case it is not what was expected.

Through the use of these learned standards, predictive monitoring in business processes aims to improve the degree of confidence of the results through different methods. This monitoring is done by combining, applying and studying the variables in the process model. The purpose of this end-of-course project was to use as a base a study on how free text combined to other attributes of process models improves prediction results and apply similar settings to another context to reinforce the hypothesis that this combination performs better than the use of only one type of information. The results show that, in general, this combination presents better performance, but also that the algorithms behave differently according to the context.

Keywords: Business Processes, Process mining, Predictive monitoring.

Sumário

Índice de Tabelas

Índice de Figuras

1. Introdução

1.1. Motivação

Há algum tempo, o tema de gestão de processos de negócio é bastante discutido e se faz presente na vida das instituições, sejam elas empresas privadas ou do setor público e junto a isso outras técnicas são abordadas como, por exemplo, a descoberta de modelos de processo, a melhoria desses modelos e o seu uso no auxílio à tomada de decisão. Com o aumento significativo do fluxo de dados e devido às informações a eles associadas, a gestão desses processos passa a apresentar um novo grau de complexidade. Como apontado por Oliveira e Bertucci (2003), a gestão da informação tornou-se um instrumento estratégico necessário para controlar e auxiliar decisões, através de melhorias no fluxo da informação, do controle, análise e consolidação da informação para os usuários.

Através do uso correto dessas técnicas, o tema de gestão de processos de negócios, apesar de mais complexo, se torna mais completo e abre caminho para outras atividades como, por exemplo, a predição de resultados da execução desses processos. Com o uso de bases de dados mais robustas e abrangentes, conseguimos analisar atributos associados às atividades do processo que não eram levados em consideração, mas que por sua vez, também influenciam no resultado ao final da sua execução. Junto a esses benefícios surgem alguns desafios. Um exemplo desses desafios é o alto consumo de tempo no processamento dos algoritmos usados para análise, como mostrado por Teineema et al. (2016) em seu artigo.

Diversas são as áreas de aplicação das técnicas presentes na gestão de processos de negócio, como abordado por Teinemaa et al. (2016) que, com seu estudo sobre a predição em processos de negócio, faz uso de técnicas de mineração de textos associados ao monitoramento preditivo para prever possíveis resultados do processo de contrato de empréstimo de uma agência bancária. Ao coletar informações como a renda e bens em posse de um contratante e informações pessoais como idade, tempo e tipo de emprego, analisar essas mesmas informações e comparar ao histórico existente, provenientes do histórico de casos similares anteriores, é possível apontar futuros atrasos e não

pagamentos por parte do cliente e sendo assim, a agência pode repensar o tipo de contrato a ser firmado, bem como negar o valor pedido pelo cliente.

Maggi et al. (2013) abordam outro tema, saindo do uso focado nos processos empresariais e se voltam para como a gestão correta dos processos de atendimento de um hospital pode ajudar no diagnóstico de doenças e indicação de tratamento adequado para seus pacientes, com base em casos similares do passado. Esses, dentre outros casos, conseguem exemplificar o quão abrangente a gestão de processos de negócio e o uso de suas técnicas são.

Tendo como ponto de referência alguns dos trabalhos desenvolvidos na área e com o intuito de explorar a aplicação de técnicas de predição em processos de negócio, este projeto teve como uma de suas etapas, reproduzir um framework criado durante o estudo da predição com uso de dados estruturados e não estruturados feito por Teinemaa et al. (2016) e aplicar técnicas semelhantes no log de um processo de resolução de incidentes para verificar como configurações semelhantes a esse framework se comportam em um outro contexto.

1.2. Objetivos

O objetivo deste projeto foi analisar a influência que as mensagens de texto ligadas às atividades de um processo de negócio têm no resultado final da execução desse processo e reforçar que a informação extraída dessas mensagens, combinada aos outros atributos presentes no processo, melhoram a eficiência da predição desse resultado.

O contexto foi dado por um processo de andamento de chamados e a meta foi utilizar as mensagens de texto trocadas entre as pessoas durante a execução de cada chamado para prever de maneira mais eficaz se um determinado caso em execução irá ultrapassar ou não o tempo de resolução pré-estabelecido ao final do atendimento. Outras análises sobre o comportamento dos diferentes métodos de predição em processos de negócios

foram realizadas durante o projeto, uma vez que possuem ligação direta com os resultados apresentados.

1.3. Organização do texto

O presente trabalho está estruturado em capítulos e, além desta introdução, será desenvolvido da seguinte forma:

- Capítulo II: Fundamentação Teórica – Apresentação dos assuntos e técnicas necessários para o bom entendimento do que foi desenvolvido e estudado neste projeto.

- Capítulo III: Trabalhos Relacionados – Exemplos do uso da mineração preditiva com foco na resolução de problemas e auxílio na tomada de decisão.

- Capítulo IV: Desenvolvimento do projeto – Aplicação e análise das técnicas de predição em processos de negócio estudas.

- Capítulo V: Conclusões – Reúne as considerações finais, assinala as contribuições da pesquisa e sugere possibilidades de aprofundamento posterior.

2. Fundamentação teórica

Este projeto tem seu foco voltado para a aplicação de técnicas de mineração preditiva de processos de negócio e como o uso dos atributos não-estruturados podem melhorar sua performance. Como mencionado na introdução do documento, diversas são as áreas de aplicação para o tema e, por consequência. Outros assuntos são discutidos e estudados em paralelo.

Com novas tecnologias emergindo a todo instante, os temas voltados para predição em processos de negócio se combinam com o objetivo de fornecer uma gestão aprimorada de informação. A analogia é com a mineração de dados que, por exemplo, facilitou o processo de encontrar correlações dentro dos grandes bancos de dados. Essas técnicas, porém, não têm seu foco em processos e são usadas para extrair informações do determinado contexto, mas sem levar em conta a ordem de execução. Já a mineração de processos concentra-se em processos de ponta a ponta e é possível devido à crescente disponibilidade de dados de eventos e novas técnicas de descoberta de processos e verificação de conformidade [Aalst, 2012].

Os temas de mineração de dados e mineração de processos são alguns exemplos dos diferentes conceitos que possibilitam a predição em áreas distintas. Neste capítulo, serão abordados esses e outros temas necessários para o entendimento do projeto e seus resultados.

2.1. Mineração de dados

Mineração de dados (MD), ou *Data Mining* em inglês, é definida como o processo de extrair informações das grandes bases de dados, para descobrir conexões ocultas e prever tendências. Muitas vezes ela será referida como "descoberta de conhecimento em bancos de dados".

MD se relaciona diretamente com inteligência artificial (IA) e com o uso de aprendizado de máquina, que é definido pela aplicação de algoritmos que aprendem com registros de dados para prever padrões. Essa combinação torna possível a implementação e melhoria de softwares que possam aprender autonomamente. Em outras palavras, um

dos objetivos de MD é fazer com que os computadores atuem sem serem explicitamente programados e possam aprender com os dados do passado, junto com aqueles gerados em tempo de execução. Novas tecnologias, como carros autônomos, reconhecimento de fala e pesquisa eficaz na web, existem devido ao aumento significativo de estudos e descobertas dentro do tema.

Um dos princípios chave para MD é a identificação de padrões de comportamento dentro das bases de dados estudadas. Para Bishop (2006), "*o campo do reconhecimento de padrões preocupa-se com a descoberta automática de regularidades em dados através do uso de algoritmos de computador e com o uso dessas regularidades para realizar ações como classificar os dados em diferentes categorias.*"

O esquema ilustrado na Figura 1 mostra como é feita a abordagem de MD para extração de informação. Na maioria das vezes, as bases de dados apresentam informações extras no registro que podem atrapalhar o processo extração de informação e que devem ser retiradas da base antes da análise. O segundo passo é colocar essa base de dados na estrutura interpretada pelo método ou algoritmo utilizado (explicados ao longo deste capítulo) para que possa servir como entrada para aprendizagem de um modelo. Somente após esses passos é que será feita a extração de padrões para análise.

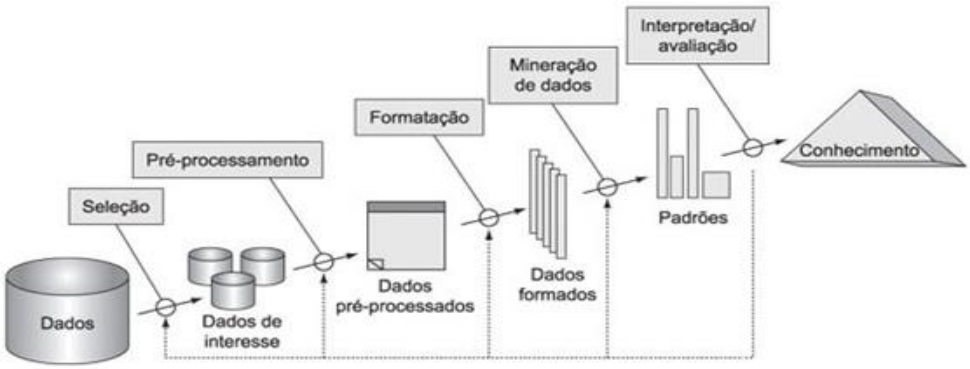


Figura 1 - Uma visão geral das etapas que compõem o processo de descoberta de conhecimento em banco de dados [Fayyad et al., 1996].

Ponto chave para extração de informação, as bases de dados podem ser representadas de inúmeras formas. O método mais comum é o de informações estruturadas em tabelas, onde as colunas definem os tipos de dados e a linhas, os valores

associados a cada uma de suas entradas. A Tabela 1 mostra uma estrutura para o registro dos dados de pacientes e seus respectivos diagnósticos.

Tabela 1 - Cadastro de pacientes e seus respectivos diagnósticos [Carvalho et al., 2012].

Idade	Sexo	Peso	Altura	Estado Civil	Profissão	Jornada	Intervalo	Horas	Dieta	Atividade física	Diagnóstico
12	F	40	1,58	Solteiro	Estudante	0	Sim	2	?	Sim	Cifose
15	F	50	1,57	Solteiro	Estudante	0	Sim	2	Não	Não	Escoliose
32	F	71	1,65	Casado	Auxiliar de limpeza	8	Sim	1	?	?	Escoliose
14	M	57	1,72	Solteiro	Estudante	0	Sim	2	?	?	Escoliose
31	M	61	1,71	Casado	Técnico eletrônico	8	Sim	1	Não	Sim	Cervicobraquialgia
36	M	90	1,83	Solteiro	Chaveiro	11	Não	0	Não	Sim	Fratura
20	M	78	1,69	Solteiro	Estudante	0	Sim	2	?	Sim	Lesão
50	M	88	1,70	Casado	Motorista	12	Sim	2	Não	Sim	Lesão
43	F	58	1,70	Divorciado	Representante comercial	44	Sim	2	?	?	Lesão
31	M	89	1,84	Casado	Representante comercial	8	Sim	1	Sim	Sim	Lesão
22	M	85	1,75	Solteiro	Estudante	0	Sim	2	Não	Sim	Lesão
45	M	78	1,76	Casado	Técnico em manutenção	?	Sim	2	Não	Não	Hérnia de disco
24	M	79	1,70	?	Atleta	?	Sim	2	?	Sim	Osteíte púbica
76	F	63	1,58	Viúvo	Do lar	44	Não	0	Não	Não	Tendinose
70	M	88	1,74	Casado	Professor	4	Sim	2	Não	Não	Ruptura de tendão do quadríceps
11	M	60	1,59	Solteiro	Estudante	0	Sim	2	Não	Sim	Fratura

Nesse caso, as técnicas de mineração poderiam ser aplicadas à base de dados para encontrar a relação entre suas colunas (diferentes tipos de dados do registro) e descobrir as causas do diagnóstico recebido pelo paciente, por exemplo. Outras relações podem ser extraídas do mesmo grupo de dados, mas é importante que se tenha definido o objeto, ou atributo que se deseja usar como a classe resultante dessa relação.

Em outro exemplo, Tabela 2, vemos os dados referentes às condições climáticas para a prática de um esporte determinado.

Tabela 2 - Condições climáticas para jogo [Castro, 2008].

Tempo	Temperatura	Umidade	Vento	Jogar
Ensolarado	Quente	Alta	Não	Não
Ensolarado	Quente	Alta	Sim	Não
Fechado	Quente	Alta	Não	Sim
Chuvoso	Branda	Alta	Não	Sim
Chuvoso	Fria	Normal	Não	Sim
Chuvoso	Fria	Normal	Sim	Não
Fechado	Fria	Normal	Sim	Sim
Ensolarado	Branda	Alta	Não	Não
Ensolarado	Fria	Normal	Não	Sim
Chuvoso	Branda	Normal	Não	Sim
Ensolarado	Branda	Normal	Sim	Sim
Fechado	Branda	Alta	Sim	Sim
Fechado	Quente	Normal	Não	Sim
Chuvoso	Branda	Alta	Sim	Não

Aqui, se definirmos jogar como a classe que se deseja prever, podemos enxergar uma relação de como as condições do tempo irão influenciar no seu valor. Em uma rápida análise para os casos apresentados na tabela, podemos dizer que se o dia estiver ensolarado e estiver quente, não haverá jogo, pois em nenhum dos casos em que essas informações são verdadeiras, o valor para a classe “Jogar” é “Sim”.

Pode-se observar que, mesmo apresentando estruturas relativamente simples, a visualização da relação entre as classes pode não ser trivial. O mesmo acontece para descobrir qual das classes é influenciada pelos valores das outras e pode ser utilizada para representar a relação dessas com o modelo avaliado. Com o volume gigantesco de informações geradas na era digital, as bases dados usadas muitas vezes apresentam proporções onde é impossível realizar essas extrações manualmente.

A solução para o processamento de grandes volumes de informações é dada pelo uso dos métodos de aprendizagem e seus algoritmos, que realizam esse processamento de maneira automatizada. O método inserido dentro do framework aqui estudado é o de Classificação, definido como aprender uma função que mapeia, ou classifica, um item de dados em uma das várias classes predefinidas por seu algoritmo [Weiss e Kulikowski, 1991]. A Figura 2 mostra a divisão de dados referentes às aplicações de empréstimos e

seus resultados (‘x’, para aceitos e ‘o’, para rejeitados), em duas classes. Em cinza, assume-se o padrão para os casos indicados como rejeitados e em branco, os casos concedidos. Nessa análise, são considerados o valor total solicitado e o salário do requerente. Note que existirão casos erroneamente classificados.

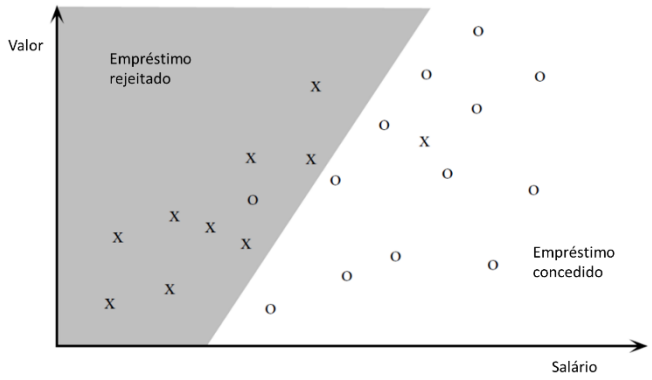


Figura 2 - Índice de empréstimos aprovados.

Muitas são as tarefas, as aplicações e as abordagens dos algoritmos de MD existentes. Suas características devem ser estudadas antes do seu uso, pois cada um deles irá apresentar resultados mais ou menos significativos, de acordo com o tipo de dado utilizado para treinamento e teste. Neste projeto, são utilizados dois modelos de classificação, Florestas Aleatórias e Regressão Logística, por apresentarem bons resultados em várias configurações de problemas e por se adaptarem bem ao caso em que os dados são muito escassos [Teinemaa et al., 2016]. Ambos são explicados a seguir.

Florestas Aleatórias (FA), ou *Random Forests* – parte da ideia de se integrar diferentes metodologias em um só modelo preditivo. Métodos em conjunto podem ser usados para melhoria em predição [Rokach, 2010]. FA pode ser inserido nesse conceito, pois é baseado na junção de árvores de decisão, definidas como explicado a seguir.

Para entendimento do algoritmo de FA é importante entender o conceito de árvores de decisão, que se trata de um modelo usado em tarefas de classificação e que possui árvores como estrutura de dados subjacentes formadas por um conjunto de elementos que armazenam informações chamadas nós. Toda árvore possui um nó raiz, que é considerado o ponto de partida, apresenta maior nível hierárquico e possui ligações com outros nós, considerados filhos. No nível mais baixo de uma árvore, estão os nós que

não possuem filhos e esses são chamados de nós folha ou terminais. Em uma árvore de decisão, uma decisão é tomada através do caminho que se inicia no nó raiz e vai até o nó folha (decisão de acordo com as escolhas). A Figura 3 mostra um exemplo de árvore para a decisão a ser feita de acordo com as condições climáticas do dia.

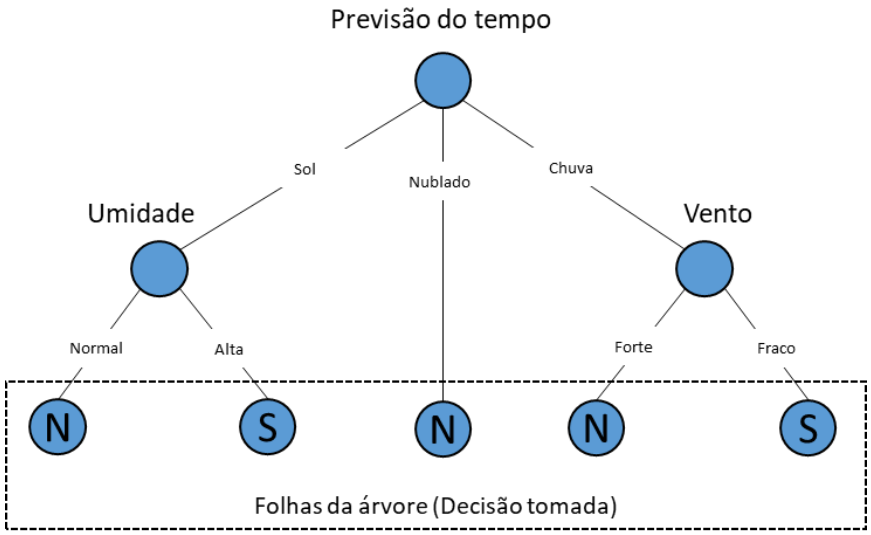


Figura 3 - Exemplo de árvore de decisão.

Na etapa de treinamento, a árvore de decisão é contruída a partir de um conjunto de treino com exemplos previamente classificados e, posteriormente, na etapa de avaliação, outros exemplos serão classificados de acordo com essa mesma árvore. A árvore apresentada acima, por exemplo, poderia ser retirada de uma tabela que mostra as condições climáticas para diferentes dias e diz, para cada um desses dias, se foi possível a prática de determinado esporte.

Existem diferentes abordagens para aprendizado e construção de árvores de decisão. Como aponta Quilan (1986), "*Uma abordagem para a tarefa de indução acima seria gerar todas as árvores de decisão possíveis que classificam corretamente o conjunto de treinamento e selecionar o mais simples deles. O número dessas árvores é finito, mas muito grande, portanto, essa abordagem só seria viável para pequenas tarefas de indução*".

Utilizando as árvores de decisão, as FA são criadas através da definição de um parâmetro de estimadores onde cada árvore resulta em uma classificação sobre uma dada

observação. O resultado é dado pela moda estatística das categorias, usado para a classificação. A Figura 4 mostra um exemplo para construção de FA.

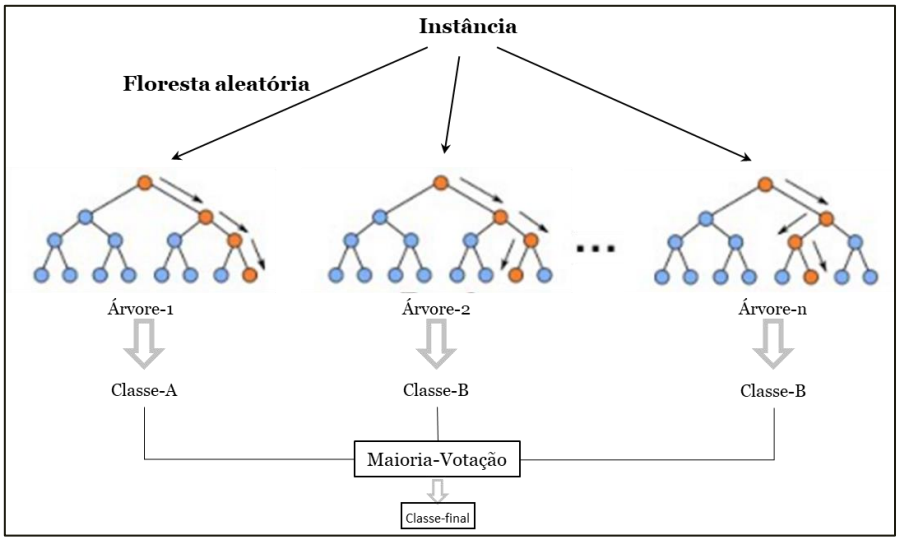


Figura 4 - Algoritmo floresta aleatória [Jagannath, 2107].

Regressão Logística (RL), ou *logistic regression* - modelo linear onde as probabilidades de resultados de uma amostra são modeladas através de uma função logística e os dois possíveis valores de variáveis dependentes são rotulados como "0" e "1", que representam resultados com valores lógicos de sim ou não e é usada para estimar a probabilidade de uma resposta binária com base em uma ou mais variáveis preditoras [Walker, 1967]. Isso permite dizer que a presença de um fator de risco aumenta as chances de um determinado resultado por um fator específico. Essa função é definida por:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Em que:

- *e*: número de Euler;
- x_0: valor de x no ponto médio da curva;
- *L*: valor máximo da curva;
- *k*: declividade da curva.

Quando representada graficamente, como na Figura 5, a função apresenta uma curva sigmoide. Em RL, o classificador modela a curva da melhor forma possível, com

base nos dados de treinamento e quando o modelo estiver ajustado, ele será avaliado com os dados de teste, diferentes da etapa de treinamento. Dessa forma, teremos as predições e podemos utilizar as métricas para calcular a precisão do modelo.

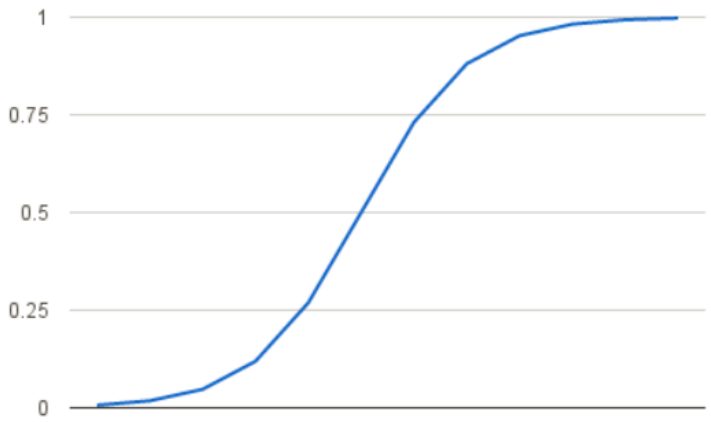


Figura 5 - Gráfico da função de regressão logística.

Apesar de muito importante e do alto valor das informações que podem ser retiradas pelo emprego das técnicas de MD, essas informações só conseguem traduzir as relações dos dados estruturados da base analisada. Dados que apresentam um formato de entrada e distribuição bem definidos, como é o caso das tabelas de bancos de dados. Porém, além desses dados estruturados, outras informações podem estar presentes na base de dados, sem possuir um padrão pré-estabelecido, como é o caso da troca de mensagens de texto durante a execução de um processo.

Em casos onde atributos não-estruturados, como texto, estão presentes, os métodos existentes de MD podem sofrer com a perda na qualidade da informação encontrada, uma vez que eles não realizam uma análise do contexto em que o texto está inserido para extração dos dados. Isto se aplica a esse projeto, pois durante o processo de resolução dos chamados presentes no processo estudado, existe uma troca de e-mails entre o usuário e solucionador. O corpo desses e-mails, muitas vezes representa atividades que foram ou que precisam ser realizadas, ou informações de registro do problema encontrado.

Visando solucionar esse tipo de problema, outras técnicas são combinadas à mineração de dados, como, por exemplo, a mineração de texto, tema abordado no próximo tópico deste projeto. Será apresentado também como a combinação desses dois temas é importante para a gestão de processos de negócio.

2.2. Mineração de texto

Como abordado no capítulo anterior, a mineração de dados trouxe um grande valor para a análise de grandes bases de dados e na extração das relações e dos padrões de comportamento a elas ligados. Porém, no final desse mesmo capítulo, é introduzido o desafio existente para o tópico, o processamento do grande volume de informação. O outro ponto é que sua abordagem é feita considerando-se somente o modelo com dados estruturados. Além dos dados estruturados e não estruturados, existem também os dados semiestruturados e a Tabela 3 mostra as principais características de cada um deles.

Tabela 3 - Tipos de dados.

Dados estruturados	Dados semiestruturados	Dados não estruturados
Esquema pré-definido	Nem sempre há um esquema pré-definido	Não há esquema pré-definido
Estrutura regular	Estrutura irregular	Estrutura irregular
Estrutura independente dos dados	Estrutura embutida nos dados	Pode não ter estrutura alguma
Estrutura reduzida	Estrutura extensa que depende da particularida dos dado e da organização	Estrutura extensa que depende da particularida dos dado e da organização
Fracamente evolutiva	Fortemente evolutiva (normalmente apresenta alto nivel de modificações)	Fortemente evolutiva (normalmente apresenta alto nivel de modificações)
Precistiva (esquemas fechados e restrições de integridade)	Estrutura descritiva	Estrutura descritiva
Distinção entre estutura e dados é clara	Distinção entre estrutura e dados não é clara	Distinção entre estrutura e dados não é clara

Na Figura 6, é mostrado o conteúdo do texto normalmente presente no corpo de e-mails. Para esse tipo de registro, existe uma peculiaridade, que é a possibilidade de ser tratado de duas maneiras. Ele pode ser considerado um tipo de dado semiestruturado, quando se consideram os campos (remetente, assunto, destinatário, etc.) existentes, devido ao fato desses campos apresentarem um formato pré-definido, mas que, ao mesmo tempo, não apresentam um padrão quando analisados separadamente. Por

outro lado, quando se deseja analisar somente o texto do corpo do e-mail, por exemplo, ele será do tipo não estruturado. Neste projeto, se aplica o segundo caso.

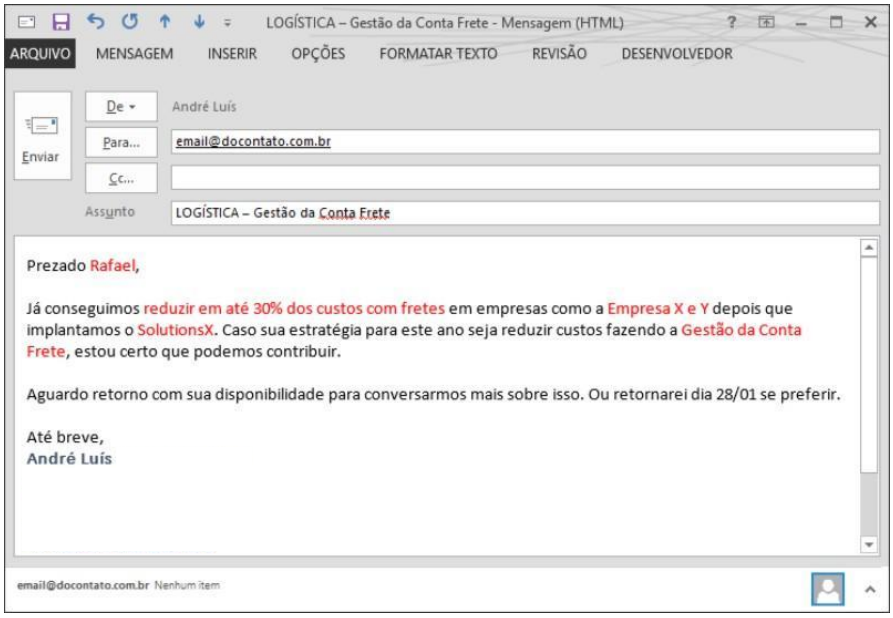


Figura 6 - Exemplo de corpo de e-mail.

A extração de conhecimento de documentos de texto (Mineração de Texto, ou MT) é um dos assuntos mais abordados hoje em dia dentro da ciência da computação, devido ao volume desse tipo de informação circulando na internet, ambientes empresariais e pesquisa. Esses dados podem ser registrados através da interação entre os próprios usuários, como a troca de mensagens por e-mail ou registro das conversas em um aplicativo, ou no registro de informação para o uso de sistemas, como registro de um problema encontrado ou o detalhamento de uma configuração.

O conceito central para a mineração de texto é o documento que, segundo Feldman e Sanger (2002), pode ser definido como uma unidade de dados textuais discretos dentro de uma coleção que normalmente, mas não necessariamente, correlaciona-se com algum documento da vida real como relatório de negócio, memorando legal, e-mail, trabalho de pesquisa, artigo, etc. As Figuras 7, 8 e 9 exemplificam alguns desses documentos.

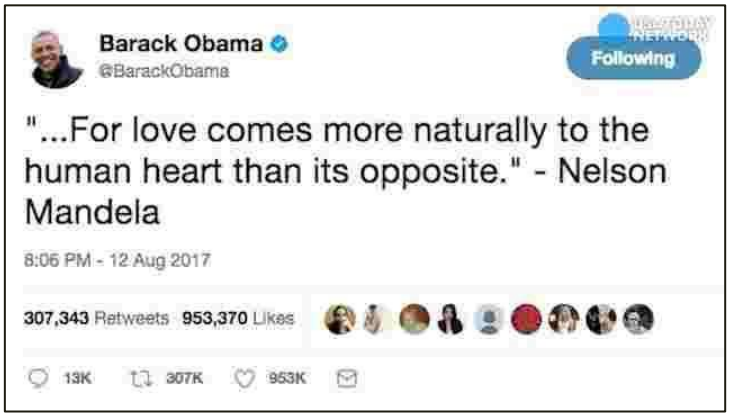


Figura 7 - Exemplo de mensagem no Twitter (rede social).

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE PELOTAS
PRÓ-REITORIA ADMINISTRATIVA
COORDENAÇÃO DE MATERIAL E PATRIMÔNIO

Memorando Cirular nº 010/2016/CMP Pelotas, 04 de abril de 2016.

Ao Gabinete Do Reitor, Pró-Reitores, Diretores De Unidades Ac adêmicas Ou Administrativas

Assunto: **Rotina para solicitação de Notificação a Fornecedores.**

Prezados(as) Senhores(as),

A fim de dar maior celeridade e eficiência aos processos de aplicação de penalidade, atendendo ao art. 87 da Lei 8.666/1993, bem como promover a redução dos Restos a Pagar, informamos que as unidades devem, assim que verificar atraso ou discordância na entrega do material solicitado, preencher o formulário **SOLICITAÇÃO DE NOTIFICAÇÃO A FORNECEDORES** disponível da página da PRA em http://wp.ufpel.edu.br/pra/cmp/material/solicitacao-de-notificacao-a-fornecedores/ , e encaminhar ao Núcleo de Material para o e-mail pendencias.ufpel@gmail.com .

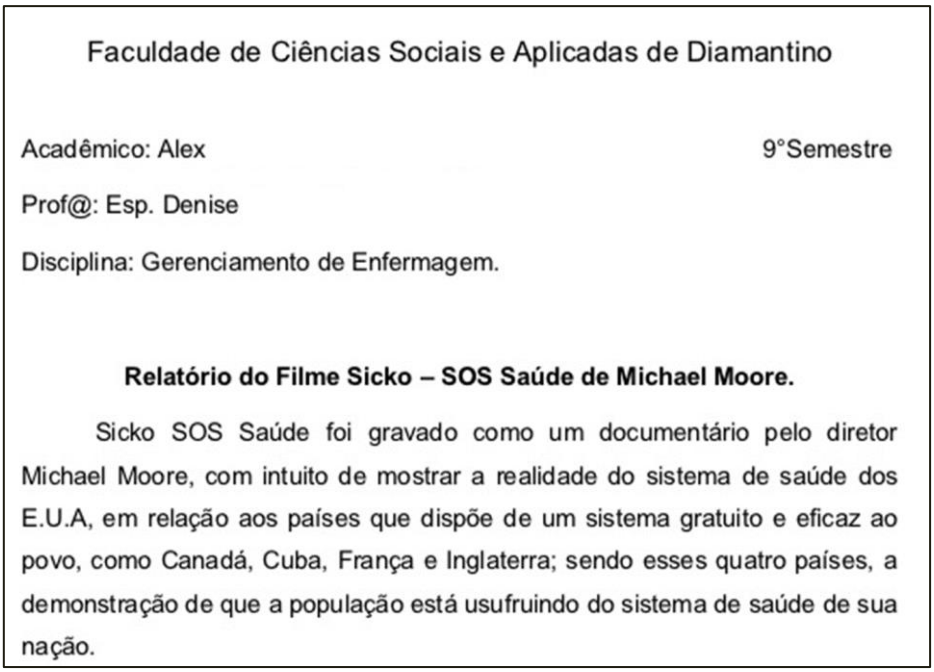
Ressaltamos a importância de tal procedimento, considerando que as informações contidas na solicitação permitirá a adequada apuração de infrações administrativas das empresas.

Atenciosamente,

Eliara Santos da Silva
Coordenadora da CMP

Antonio Carlos de Freitas Cleff
Pró-Reitor Administrativo

Figura 8 - Exemplo de documento administrativo.



Faculdade de Ciências Sociais e Aplicadas de Diamantino

Acadêmico: Alex 9°Semestre

Prof@: Esp. Denise

Disciplina: Gerenciamento de Enfermagem.

Relatório do Filme Sicko – SOS Saúde de Michael Moore.

Sicko SOS Saúde foi gravado como um documentário pelo diretor Michael Moore, com intuito de mostrar a realidade do sistema de saúde dos E.U.A, em relação aos países que dispõe de um sistema gratuito e eficaz ao povo, como Canadá, Cuba, França e Inglaterra; sendo esses quatro países, a demonstração de que a população está usufruindo do sistema de saúde de sua nação.

Figura 9 - Exemplo de relatório de trabalho.

Podemos ver através dos exemplos que cada tipo de texto possui um padrão diferente de escrita. O *tweet*, por exemplo, por ter o número de caracteres reduzido para cada comentário, possui frases curtas e muitas vezes suas palavras são cortadas em pedaços para caberem no tamanho definido. As pessoas que fazem uso frequente da rede social estão cientes dessa condição e mesmo termos que não pertencem ao vocabulário formal da língua são entendidos pelas pessoas inseridas no contexto. Já os outros exemplos de texto, apresentam vocabulário específico para seus assuntos. Todos os textos, porém, passam uma mensagem e devem ser entendidos durante a interação com seus receptores, resta então conseguir extrair seus padrões, para que seja computacionalmente viável o processamento, com menor perda possível de valor.

Sendo assim, o objetivo da MT é o de se extrair padrões não-triviais e relevantes ao contexto ou processo a que o documento está relacionado. A ideia é coletar esses dados, na maior parte das vezes não estruturados, e torná-los disponíveis para análise. Isso é possível combinando técnicas de Mineração de Dados, Aprendizado de Máquina, Processamento de Linguagem Natural (NLP), Recuperação de Informação (IR) e Gestão do Conhecimento.

Devido à sobrecarga de informação disponível, hoje estima-se que a mineração de texto apresenta um apelo comercial maior do que o para mineração de dados, o que é facilmente visualizado quando estudos mostram que a maior parte das informações de uma instituição, cerca de 80%, é armazenada em documentos de texto [Tan, A. H. 1999]. Diversas são as empresas que oferecem sistemas destinados à mineração e processamento de textos para extração de informação.

Um ponto importante para o assunto é que, mesmo possuindo um alto valor comercial oriundo do potencial informativo que esses documentos apresentam, o processamento para tornar possível a extração de padrões não-triviais de documentos de textos, também apresenta uma complexidade muito maior, quando comparado ao processamento para retirar esse tipo padrão de base de dados estruturadas. Para exemplificar, podemos citar os desafios relacionados à dimensionalidade dos documentos, uma vez que as palavras ali inseridas podem ser combinadas de muitas formas diferentes.

Genericamente falando, a mineração de texto pode ser dividida em três etapas, como ilustras a Figura 10.

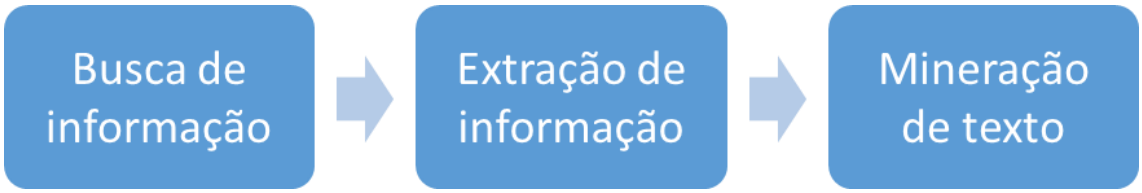


Figura 10 - Etapas da mineração de texto.

Busca de Informação – A primeira etapa é a busca de informações, onde é selecionado o corpo dos documentos de texto a serem utilizados, que será colocado no formato padrão.

Extração de informação – A segunda etapa é a marcação do texto, com o objetivo de identificar seu sentido e extrair informações automaticamente, e muitas são as abordagens para a extração de informação. Uma dessas abordagens é baseada em dicionários, onde as palavras que o compõem são selecionadas de acordo com o diagnóstico e o tipo de informação importante do contexto. Uma comparação dessas

palavras (ou frases) presentes nos documentos é feita, indicando qual a relação delas com o tema.

Mineração de Texto – Que, como explicado anteriormente, é a aplicação dos algoritmos para extração de padrões do texto analisado.

Existem diversas técnicas utilizadas para mineração de texto, como o método *bag-of-n-grams* (BoNG), que parte dos mesmos princípios do modelo *bag-of-words* (BoW), ou saco-de-palavras, e tem o texto representado por uma coleção de palavras que independem da gramática. Esse grupo de palavras pode ser formado separando cada termo independentemente e mostrando a frequência com que eles aparecem. Como exemplo, a frase "Não posso falar agora, te ligo depois", poderia ser representada por {"não":1, "falar":1, "agora":1, "te":1, "ligo":1, "depois":1}. Neste caso, porém, não é possível representar a ordem em que essas palavras estão dispostas no documento. O modelo BoNG soluciona esse problema, uma vez que a seleção das palavras é feita em grupos de n-palavras. A representação do mesmo documento em grupos de duas palavras ficaria então como a seguir {"não posso":1, "posso falar":1, "falar agora":1, ...}. Outras atividades podem ser aplicadas durante esse processo como a classificação e lematização, ambas detalhadas por [Feldman e Sanger, 2002], e devem ser escolhidas de acordo com o contexto estudado. Outras abordagens e suas variações são detalhadas por [Blei et al. 2003].

A extração de informação de alta qualidade de textos é aplicada a diversos temas como a melhoria dos mecanismos de busca existentes na internet, ou para resolução de problemas dentro da biomedicina como a melhoria e predição de seus processos de análise clínica. Nesse projeto, abordamos como as técnicas de MD e MT e suas variáveis aqui descritas, combinadas à mineração de processo, podem ajudar dentro do tema de inteligência de negócio, mostrando seu impacto na melhoria do monitoramento preditivo. A mineração de processos e suas características são explicadas dentro do próximo tópico deste documento.

2.3. Mineração de processo

Os modelos de processos são utilizados com intuito de definir e documentar como as atividades devem ser executadas. Como explica Dumas et al. (2013), "*a maneira em que os processos são estruturados e executados afeta tanto a qualidade do serviço percebida pelos clientes, quanto a eficiência em que esses serviços são entregues e uma organização pode superar outra oferecendo os mesmos tipos de serviços, desde que tenham processos melhores e os executem da forma adequada*". Através de documentos e de uma estrutura de gerência de processos bem definida, é possível realizar o monitoramento, garantir que as etapas dos processos sejam realizadas da forma mais eficaz e que o recurso e tempo necessários sejam direcionados para cada uma delas.

Contudo, devido ao fluxo crescente de informação e ao grande número de pessoas e áreas envolvidas, criar e monitorar os processos de uma organização se tornam atividades com alto grau de complexidade e consumo de tempo elevado. Devido a esse aspecto, a mineração de processos, uma das diversas iniciativas dentro da gestão de processos de negócio, é utilizada por um número cada vez maior de instituições, e está presente em muitos dos trabalhos de cientistas de dados e pesquisadores, no meio acadêmico.

A mineração de processos é feita baseando-se no histórico de atividades realizadas anteriormente, ou o chamado log de eventos de um processo, que é o conjunto de todos os traços registrados em um determinado período, onde os traços são as ações realizadas individualmente pelos trabalhadores durante a execução de um processo. Assim como apresentado durante o tópico de mineração de dados, os logs de eventos também podem ser representados por estruturas semelhantes, como mostra a Tabela 4, a diferença é que cada linha aqui irá representar as atividades executadas durante a instância do processo e os atributos relacionados à cada uma delas.

Tabela 4 - Exemplo de um log de eventos.

ID do Caso	Hora de início	Hora de término	Atividade	Ator
258	16/02/2017 12:05:00	16/02/2017 12:25:00	Criar pedido de compras	Solicitante
258	16/02/2017 13:15:00	16/02/2017 13:20:00	Analisar pedido de compras	Responsável pelo solicitante
258	16/02/2017 13:22:00	16/02/2017 13:23:00	Solicitar cotação para pedido de compras	Responsável pelo solicitante
258	17/02/2017 09:00:00	17/02/2017 11:05:00	Analisar solicitação de cotação	Responsável de compras
258	17/02/2017 11:08:00	17/02/2017 11:09:00	Aprovar cotação	Responsável de compras
258	17/02/2017 15:10:00	16/02/2017 15:45:00	Confirmar compra para o pedido	Solicitante
386	27/07/2017 10:30:00	27/07/2017 11:17:00	Criar pedido de compras	Solicitante
386	27/07/2017 17:08:00	27/07/2017 18:05:00	Adicionar informações ao pedido de compras	Solicitante
386	28/07/2017 09:15:00	27/07/2017 10:05:00	Analisar pedido de compras	Responsável pelo solicitante
386	28/07/2017 10:10:00	28/07/2017 10:12:00	Solicitar cotação para pedido de compras	Responsável pelo solicitante
386	29/07/2017 09:18:00	29/07/2017 10:00:00	Aprovar cotação	Responsável de compras
386	29/07/2017 10:02:00	29/07/2017 10:06:00	Encaminhar cotação para fornecedor	Responsável de compras
386	29/07/2017 11:50:00	29/07/2017 12:23:00	Selecionar melhor opção	Fornecedor
386	29/07/2017 17:17:00	29/07/2017 17:55:00	Confirmar compra para o pedido	Solicitante

Através do uso de algoritmos e ferramentas de mineração, como o ProM[1] e o Disco[2], é possível extrair os diferentes comportamentos presentes no log de eventos e defini-los em modelos que se adequem melhor ao caso analisado. Dessa forma, também é possível encontrar os desvios existentes, pois mesmo que já exista um modelo a ser seguido, durante a execução, é muito provável que haja atividades fugindo em parte ou em sua totalidade do escopo definido inicialmente, como por exemplo uma troca de responsabilidades entre recursos.

Segundo Aalst (2011) "*a mineração de processos está preenchendo a lacuna entre a análise clássica do modelo de processo e a análise orientada a dados, como mineração de dados e aprendizado de máquina. Pois além de ter seu foco voltado para os processos, diferentemente da mineração de dados clássica, faz uso de dados reais*".

A Figura 11 mostra três dos principais usos da mineração de processos:

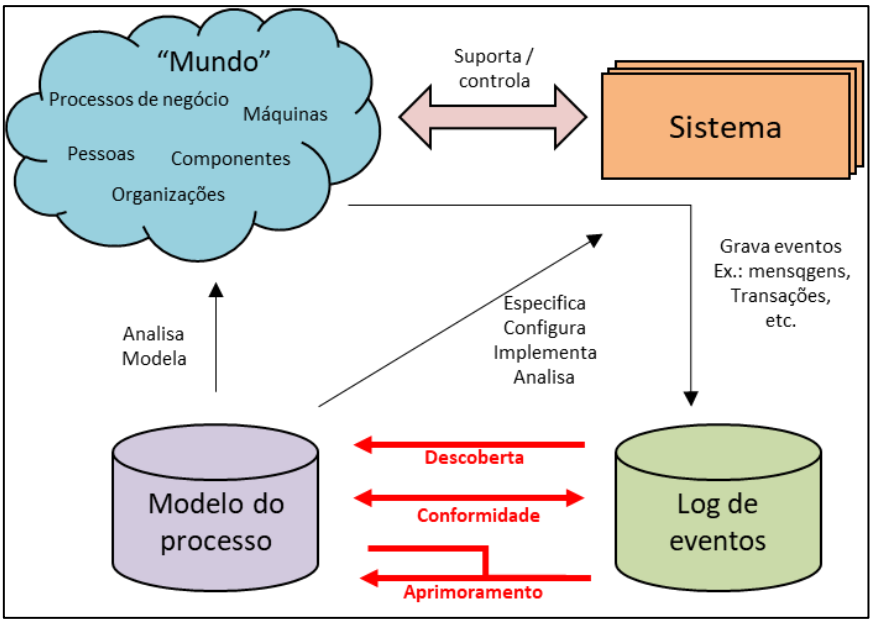


Figura 11 - Os três tipos básicos de mineração de processo: (a) descoberta, (b) conformidade e (c) aprimoramento [Aalst, 2011].

- **Descoberta** – conjunto de técnicas usadas para construir uma representação dos processos de negócio e suas principais variações, podendo ser dada de forma automática ou manual. Essas técnicas usam evidências encontradas nos métodos de trabalho existentes, nas documentações e nos sistemas dentro de uma organização.
- **Conformidade** – técnicas aplicadas ao modelo aprendido anteriormente, para compará-lo com seu log de eventos e verificar se a execução dos casos está de acordo com a representação e vice-versa. Um exemplo de como isso pode ser feito, é verificando se todas as atividades das instâncias podem ser executadas dentro daquela representação.
- **Aprimoramento** – atividade proativa de identificar, analisar e aperfeiçoar os processos de negócios existentes, visando otimizar e atender novas definições ou padrões de qualidade.

A mineração de processos pode tratar dois tipos básicos de processos: estruturados e não estruturados. Os estruturados são regulares, repetitivos e podem ser controlados, o que significa que as instâncias dos processos apresentam uma certa regularidade em relação ao seu tempo e ritmo de execução. O resultado disso é que o caminho principal do fluxo do processo estruturado normalmente é representado por um modelo

relativamente simples. Já os processos não estruturados se apresentam de modo irregular e seus eventos normalmente são flexíveis e variáveis e, por consequência, descobrir o comportamento padrão é uma tarefa complexa.

Aliado às técnicas de melhoria, está o monitoramento preditivo, que através do uso de frameworks e algoritmos, auxiliam para que seja possível prever, com alto grau de confiança, o resultado de cada instância dos processos em execução que se deseja analisar. Isso faz com que as ações necessárias para alteração do resultado previsto, como a troca de uma atividade, ou até mesmo a interrupção do processo, possam ser tomadas de forma proativa e não mais reativa. O monitoramento preditivo é mais um tema voltado para a melhoria contínua na gestão de processos de negócio.

2.4. Monitoramento preditivo

O monitoramento de processos de negócio se refere à análise dos eventos produzidos durante a execução de um processo, com o intuito de avaliar o cumprimento de seus requisitos e sua performance [Dumas et al., 2013]. Duas são as abordagens: (i) de modo *offline*, onde usa-se as informações de registros das instâncias finalizadas e (ii) online, para analisar a performance do caso em execução.

A gestão de processos de negócio e sua eficácia no auxílio à tomada de decisão está diretamente relacionada ao quanto é possível reduzir o nível de incerteza do resultado de cada uma das instâncias executadas através da predição, ou seja, quanto mais rápido e com maior nível de confiabilidade se pode prever os resultados de um processo, melhores decisões serão tomadas e mais rapidamente. Utilizando uma restrição do negócio (*business constraint*), definida como um requisito imposto à execução de um processo que separa o comportamento compatível com o não-conforme [Pesic e Aalst. 2006], o responsável pela gestão do processo irá tomar as medidas necessárias para que se alcance o resultado esperado.

A predição de processo pode ser feita de diversas maneiras, dentre elas estão as técnicas de aprendizado de máquina e mineração de dados, onde cada previsão é alimentada por dados coletados em determinado período (log de eventos de instâncias finalizadas anteriormente) que são utilizados para prever o futuro em diferentes cenários.

Com o monitoramento preditivo, é possível antecipar diversas informações durante a execução do processo como o tempo necessário para concluir uma tarefa, uma atividade ou operação específica. O grande ganho disso é o suporte para que as instituições consigam atingir suas metas previstas para cada um dos casos, podendo interromper a execução, caso vejam que ela irá sair do esperado. Outros temas também são abordados no estudo de monitoramento preditivo em processos de negócio, como a estimativa de tempo, seja de violação de prazo (SLAs) ou previsão do início e fim das tarefas, ou os casos utilizados para prever os riscos, como exemplificado por Conforti et al. (2016).

O método base utilizado para predição em processos de negócio, ilustrado na Figura 12, é dado da seguinte maneira:

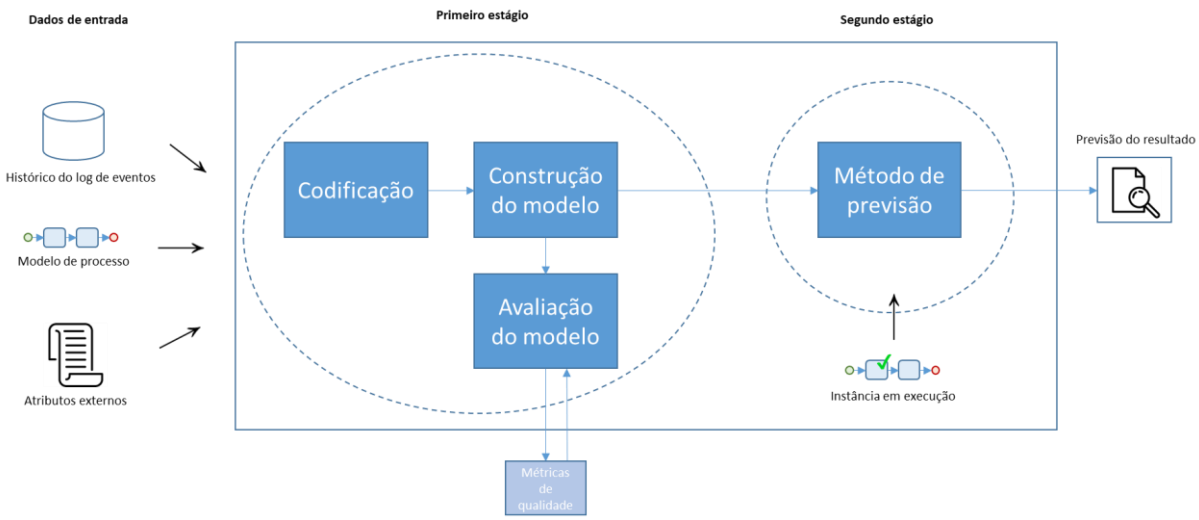


Figura 12 - Procedimento experimental de um método genérico de monitoramento preditivo [Marquez-Chamorro et al., 2017].

1. Primeiro estágio (*Offline*): onde o componente offline é utilizado para treino, seguindo as métricas definidas. Para isso, ele irá consumir informações como o modelo e os atributos do processo e os dados históricos de outras instâncias, ou o chamado log de eventos. A codificação e qualificação do modelo de treino será feita pelo uso de algoritmos voltados ao tema. Uma vez aprovado pelos critérios de qualidade, poderá ser usado na próxima etapa.

2. Segundo estágio (*Online*): usando o método de predição definido, a instância do processo em execução será avaliada verificando-se os eventos e atributos apresentados até então.

Como apontado por Aalst, (2011), três são as atividades para operação em predição:

- **Detectar**: Assim que um processo em execução se desvia do caminho mapeado no modelo, um alerta é disparado. Ponto utilizado para que o modelo existente seja atualizado com informações do que acontece na vida real durante a execução do processo.

- **Prever**: Dados históricos extraídos de logs de eventos são processados para criar modelos que ajudam a fazer previsões do resultado esperado, como tempo de conclusão, alguma classe definida como meta, entre outros.

- **Recomendar**: Com base nas previsões feitas, um modelo de recomendação propõe ações para serem tomadas para o caso analisado. Como exemplificado por Francescomarino et al. (2013), isso pode ser usado para que uma instância em andamento tenha probabilidade maior de chegar ao objetivo esperado ao final de sua execução.

Abordando o desafio apontado nos tópicos anteriores, pesquisas passam a analisar o impacto de outros tipos de informação, como o exemplo das mensagens texto trocadas durante a execução das atividades do processo, no resultado observado. Em seu artigo, Teinemaa et al. (2016) mostram que, mesmo com um alto grau de complexidade e abstração presentes nos textos, é possível melhorar consideravelmente os resultados da predição.

Com os termos apresentados, considerados necessários para o entendimento do projeto aqui desenvolvido, o próximo capitulo irá apresentar a framework desenvolvido para o monitoramento preditivo, com o uso de dados estruturados e não estruturados.

2.5. Framework para Monitoramento Preditivo de Processos de Negócios

No artigo produzido por Teinemaa et al. (2016), foi realizada uma análise sobre como o uso de texto livre, tido como atributo não estruturado do processo estudado no artigo e a sua combinação com os estruturados pode contribuir para o monitoramento preditivo de processos de negócio. A proposta apresenta um framework capaz de lidar com logs de eventos com tais características e seus resultados são comparados às técnicas existentes. Para isso, técnicas de mineração de texto foram utilizadas com intuito de extrair informações relevantes de cada caso e que no segundo momento foram aliadas às técnicas de classificação baseadas em eventos.

Para confirmar se o resultado apresenta melhorias ou não, um estudo de caso foi feito em logs de eventos de dois processos reais de uma empresa financeira: (i) processo de recuperação de dívidas, cujo resultado é o reembolso parcial do valor negociado previamente ou o encaminhamento do caso para uma agência externa de cobrança e (ii) um processo de oferta de contratos em que o resultado é a assinatura ou não de um empréstimo do cliente em potencial. O objetivo, então, é saber se o valor devido será pago no prazo definido, para o primeiro processo e se o cliente irá fechar um novo contrato, para o segundo. Em ambos os casos, essa previsão deverá ser feita em tempo de execução do processo, tendo como entrada o histórico das ações realizadas e das mensagens trocadas até o dado momento.

A primeira etapa do framework é a construção de modelos e extração das características de texto, sendo estas, o resultado da primeira extração de informação dos documentos. Essa extração é feita com base nas mensagens associadas a cada um dos eventos no registro. Diferentes técnicas e parâmetros foram utilizados para criar modelos de texto, visando verificar qual apresenta melhor resultado de acordo com as características específicas de cada log.

Na segunda etapa do framework, as características extraídas dos documentos de texto são combinadas com os atributos estruturados, ambos adicionados a um vetor de tamanho fixo e com seus valores representados por números. Isso é feito para que os

algoritmos de aprendizado de máquina consigam interpretar suas informações e assim treinar os classificadores para previsão, essa sendo a última etapa do framework.

Para criar os modelos de textos, os métodos utilizados no framework foram o BoNG, que foi detalhado no capitulo desse projeto sobre mineração de texto; Taxas de contagem de log com *Naïve-bayes* (NB), que é baseado no modelo BoNG, mas ponderado com as taxas de registro do NB, [Allahyari et al., 2017]; Alocação latente de Dirichlet (LDA, *Latent Dirichlet Allocation,* em inglês), em que o modelo de texto é representado por tópicos abordados pelos documentos; e *Paragraph Vector* (PV), em que não somente os termos, mas também suas sequências, são extraídos para criar o modelo. Os dois últimos métodos são detalhados por [Blei et al. 2003].

Como os documentos de texto presentes no processo estudado não apresentam um padrão a ser seguido, um tratamento e limpeza de dados são executados em cada um dos logs, visando manter somente informações relevantes para o contexto. Após essa etapa inicial, o framework é divido em dois componentes. O primeiro, apresentado na Figura 13 é *offline* e realiza a estruturação dos dados, combinando a sequência de ambos os tipos de atributos dos casos no histórico para treinar os classificadores que, no segundo componente, *online*, fará uso desses dados para realizar as previsões dos casos em execução.

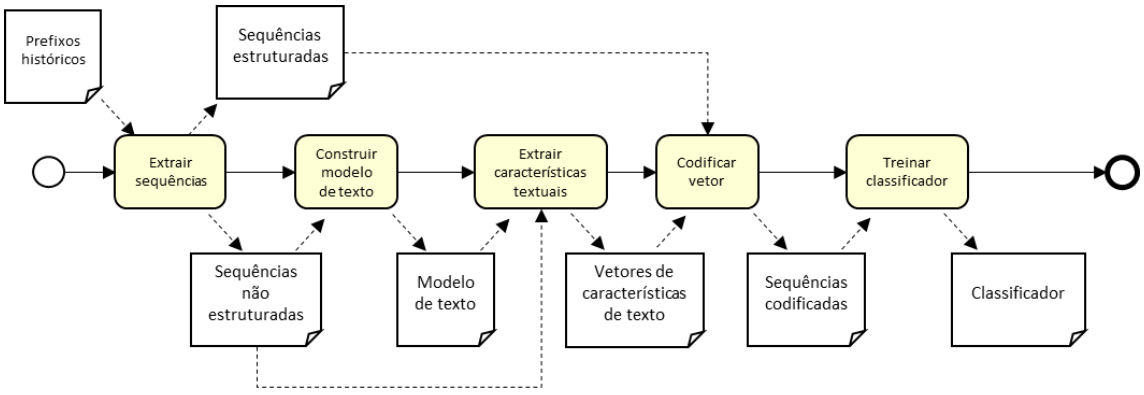


Figura 13 - Componente offline do framework [Teinemaa, 2016].

A outra parte é feita pelo componente *online* definido pelo framework, conforme a Figura 14. É nesse momento que será realizada a predição e classificação da instância em execução. Para sua execução, o grau de confiança mínimo precisa ser adicionado

como variável de entrada. Quando executado para previsão do resultado de um caso em andamento, com k eventos realizados, se a probabilidade apontada pelo classificador for maior do que esse limite, o resultado será apontado como positivo.

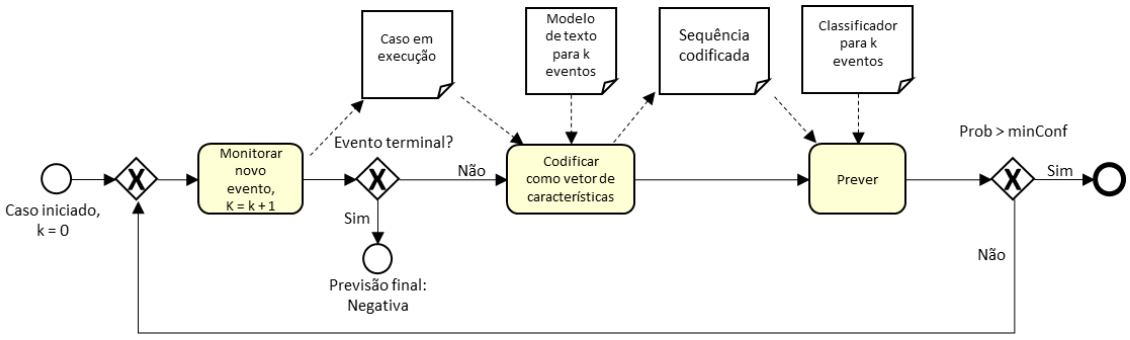


Figura 14 - Componente online do framework [Teinemaa, 2016].

Enquanto esse requisito não for alcançado, o framework continuará monitorando as próximas atividades. Se o evento observado for terminal, o resultado da previsão será classificado como negativo.

Para avalição do desempenho do framework proposto, foram propostas métricas baseadas nas combinações possíveis do valor real no registro e os resultados previstos, onde Verdadeiro Positivo (VP) e Falso Negativo (FN) são os casos originalmente positivos que foram corretamente previstos como positivos e erroneamente como negativos, respectivamente. Verdadeiro Negativo (VN) e Falso Positivo (FP), sendo originalmente negativos os que foram corretamente previstos e erroneamente previstos como positivos.

A partir daí, os valores da precisão (P), definida pelo número de casos previstos como positivos que eram realmente positivos – P = VP/(VP+FP), *recall* (R), que apresenta quais dos valores originalmente positivos foram previstos corretamente – R = VP/(VP+FN) e o *F-score* = 2*P*R/(P+R), foram usados para avaliar se os resultados apresentam uma melhora em relação aos mesmos valores dados pela *Baseline*, que é dada pelo resultado dessas mesmas variáveis, usando-se somente os atributos estruturados do processo para análise.

Os logs de ambos os processos foram divididos aleatoriamente em duas partes,

sendo uma para treinar os classificadores no componente *offline* e a segunda, menor, para testar o *online*.

O artigo informa que em uma comparação direta, os classificadores treinados com o algoritmo de florestas randômicas apresentam uma diferença expressiva se comparados aos que utilizaram regressão logística. Devido a isso, somente os resultados para o primeiro caso são exemplificados.

Na Figura 15, pode-se ver tais resultados para cada um dos métodos de criação dos modelos de texto com diferentes níveis de confiança (*minConfig*) e a comparação desses com a *Baseline* para os processos de recuperação de débito (DR) e assinatura de contrato (LtC).

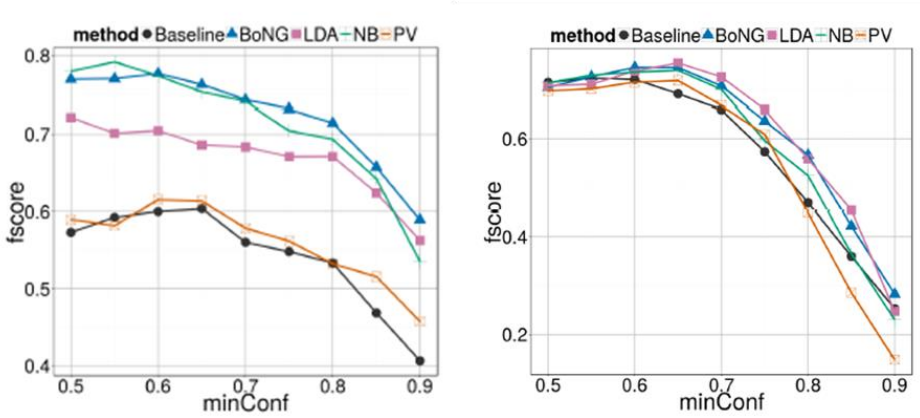


Figura 15 - Resultados para o F-score [Teinemaa, 2016].

Em ambos os casos, pode-se observar que os métodos que fazem uso das informações retiradas do texto superam quase sempre os resultados da *Baseline*. Sendo assim, já é possível afirmar que o uso de informações extraídas de texto, combinadas aos outros atributos do processo, melhora a previsão da resolução do caso em execução.

Utilizando os dados para ambos os logs, pode-se dizer também que o modelo BoNG possui o melhor desempenho, uma vez que no processo de recuperação de débito apresenta os maiores valores para o *F-score* e, no segundo caso, por mais que o modelo LDA esteja em primeiro, o ganho não é considerável. O maior valor do *F-score* obtido por BoNG foi de aproximadamente 0.78.

Por mais que o artigo analisado discorra sobre outros pontos, acima podemos observar que somente as análises relativas à precisão dos resultados foram abordadas para esse projeto. No artigo, tópicos como a brevidade que se pode realizar a previsão e o tempo necessário para rodar o framework também são abordados, porém, como a quantidade de eventos necessários para prever a variável final não é tema dessa pesquisa, os valores obtidos não afetaram as métricas utilizadas para esse projeto.

No decorrer do capítulo de desenvolvimento desse documento uma análise será apresentada comparando os resultados obtidos no artigo analisado com os apresentados neste projeto de fim de curso, de modo que seja possível visualizar como as técnicas apresentadas com melhor desempenho pela autora se comportam em uma base de dados com características diferentes. Assim é possível discorrer um pouco mais sobre as hipóteses aqui levantadas como a de que os resultados podem mudar, caso o conteúdo do texto seja mais heterogêneo.

3. Trabalhos relacionados

Aqui serão apresentadas, brevemente, outras aplicações do monitoramento preditivo em processos para a resolução de problemas. Veremos que áreas de nichos completamente diferente podem fazer uso de técnicas e conceitos muito parecidos.

Como citado anteriormente, o monitoramento preditivo (tema foco deste projeto) é um dos assuntos mais abordados dentro de BPM e, muitos trabalhos são desenvolvidos em torno do tópico. Durante o desenvolvimento do estudo e busca de informação sobre trabalhos anteriores para embasamento e definição da metodologia, uma das áreas mais presentes dentre os resultados encontrados na busca, foi a medicina. Maggi et al. (2013), por exemplo, combinam a descoberta de processos baseada no controle de fluxo de eventos com a baseada no fluxo de dados e mostram como isso poder ser usado para auxiliar nas decisões dos médicos durante uma análise clínica. Isso é feito com base nas informações de casos similares.

Outro caso de monitoramento preditivo é o apresentado por De Leoni e Aalst (2013), em que desenvolvem sobre a análise do log de eventos de um processo para lidar com empréstimos solicitados pelos clientes de um instituto de crédito.

Também inserido no nicho das organizações, sejam privadas ou não, vemos muitos trabalhos voltados para o monitoramento de riscos, como mostra Conforti et al. (2016) e (2013). Outros trabalhos possuem um olhar mais acadêmico, como o apresentado por Senderovich et al. (2016), que faz uso das redes de petri estocásticas generalizadas (GNSPs) com o objetivo de melhorar a acurácia das previsões.

Artigo base para o desenvolvimento deste projeto, criado por Teinemaa et al. (2016), mostra que, dependendo do log estudado e com o uso das técnicas corretas de classificação de texto, a previsão dos resultados do processo pode apresentar uma melhoria significativa, quando comparada às abordagens que utilizam somente dados estruturados. Além de pontos levantados durante a execução do projeto, a principal contribuição desse projeto, é a aplicação do framework em um log de diferente estrutura para reforçar ou refutar as hipóteses levantadas anteriormente.

4. Desenvolvimento do projeto

Utilizando como base os resultados do estudo apresentado em [Teinemaa et al., 2016], este projeto visa contribuir para a análise de como a combinação de dados estruturados e texto livre podem melhorar os resultados do monitoramento preditivo para processos de negócio. As técnicas que apresentaram melhor desempenho durante esse estudo, foram aplicadas a um log com características diferentes para que fosse possível expandir o conhecimento sobre as hipótese de que o uso de informações de textos associados às atividades do processo melhoram a predição, assim como trazer novas questões para estudos futuros sobre o tema.

O desenvolvimento desse projeto se dá seguindo os mesmos moldes do framework apresentado por Teinemaa em seu estudo. Isso quer dizer que as etapas principais foram executadas da seguinte maneira: 1- Extração do vetor de características do texto; 2- Combinação dos diferentes dessas características aos outros atributos do processo; e 3- Treinamento dos Classificadores.

Esse estudo realizou uma comparação entre diferentes métodos para construção do modelo de texto e extração das características de maior impacto para o processo. O modelo de texto que apresentou melhor resultado para a predição foi o *"bag of words"* e, por isso, foi o modelo escolhido para o desenvolvimento desse projeto de graduação.

O log de eventos estudado contém registros dos chamados criados pelos colaboradores de uma empresa do Rio de Janeiro quando estes encontraram problemas relacionados à tecnologia. Para resolução, um processo de atendimento desses chamados deve ser seguido de acordo com o tipo de problema encontrado. Essas informações foram registradas durante o ano de 2015 por uma ferramenta de Gerenciamento de Serviços de Tecnologia de Informação (ITSM).

O log contém informações de cada caso (ticket) e os atributos relacionados a cada um dos eventos executados durante o processo de resolução. Todos esses dados estão organizados em um documento .CSV onde as linhas representam as atividades e as colunas, os atributos atrelados a cada uma delas.

Esses atributos podem representar informações do caso, como acontece nas colunas em que o tipo de atendimento e a sua prioridade estão indicados, assim como podem mostrar o que está relacionado a cada atividade durante a execução. No primeiro caso, as informações são inseridas no momento que o processo se inicia e se manterão com o mesmo valor até o final. Já no segundo, a cada atividade, esse valor poderá ser alterado.

O atributo utilizado para análise da influência do texto no processo é o corpo dos e-mails trocados pelos solicitantes do atendimento e os membros das equipes solucionadoras. Assim como os demais, seus valores estão indicados através de uma coluna. Os valores em suas linhas podem ser nulos, o que indica que para aquela atividade, não houve troca de mensagens.

Dentro do documento, existe também um atributo que mostra para cada um dos casos se o processo foi executado dentro da SLA definida. Isso é feito através dos valores "sim" ou "não" presentes na coluna "SLA perdida". Esses valores são obtidos através da comparação do tempo de execução e o prazo pré-estabelecido para as diferentes prioridades. Caso o tempo de execução seja maior que o acordado, o valor para a coluna será "sim", indicado que o prazo foi ultrapassado.

Sendo assim, o objetivo desse projeto foi utilizar os dados de registro e a sequência das atividades de cada ticket para prever se o valor presente nessa coluna será positivo ou negativo. A seguir, estão detalhadas as etapas de execução, o log estudado e os resultados encontrados.

4.1. Definição do processo

Como mencionado anteriormente, o log utilizado apresenta registros dos casos de um processo de tratamento de tickets relacionados ao setor de tecnologia de informação para resolução de incidentes encontrados pelos usuários da mesma empresa. O objetivo desse processo é normalizar a operação, de maneira eficaz e dentro do prazo estabelecido.

O processo se inicia quando um usuário de outro setor da empresa percebe algum problema que bloqueie ou atrapalhe seu trabalho e esteja diretamente relacionado com tecnologia. Os diferentes níveis de impacto irão definir a prioridade dos chamados e consequentemente, suas SLAs.

O processo de tratamento se inicia quando um colaborador da empresa abre um chamado através da ferramenta de gerenciamento de serviços de TI, adicionando informações iniciais como, por exemplo, qual tipo de problema enfrentado. Esse ticket é então direcionado para a equipe de nível 1, que irá fazer a triagem das informações e verificar qual equipe de suporte deve ser acionada para resolução do incidente.

Caso o problema possa ser resolvido pelo próprio time, o ticket continuará na fila da equipe e as ações necessárias serão tomadas diretamente por um de seus membros. Caso o problema seja causado por uma área mais específica, o atendente irá transferir o ticket para a equipe que possui capacidade de resolução. Esse ticket poderá passar pelas demais equipes até sua resolução caso o problema esteja relacionado a mais de uma área técnica.

Após as atividades consideradas suficientes pelo responsável, o solicitante será notificado com a proposta de resolução e poderá confirmar se o problema inicial for resolvido, encerrando as atividades para o caso, ou se ainda existem problemas a serem solucionados e, nesse caso, o ticket retornará para quem propôs a solução e permanecerá nesse ciclo até que a resolução atenda aos requisitos. Quanto menos vezes o chamado precisar voltar para a equipe solucionadora, melhor será o tempo de resolução. Sendo assim, o melhor caso será aquele em que a primeira solução é aceita pelo usuário.

4.2. Log de eventos

O log de eventos utilizado possui registros da execução de casos do primeiro trimestre do ano de 2015. O arquivo contém informações referentes a 6337 tickets resolvidos durante o período e as atividades necessárias para resolução de cada um, divididos entre as 84 diferentes classificações de chamado.

Em média, a quantidade necessária de passos para a resolução do problema é de 39. Pode-se observar também os extremos em que, no menor caso, esse número foi de

somente 10 e no maior, que precisou de 357 até ser resolvido. A grande maioria dos casos não possui esse número acima de 50, como pode ser observado no gráfico da Figura 16. Em uma análise direta, observa-se que quanto maior o número de atividades executadas, maior a probabilidade de ultrapassar a SLA.

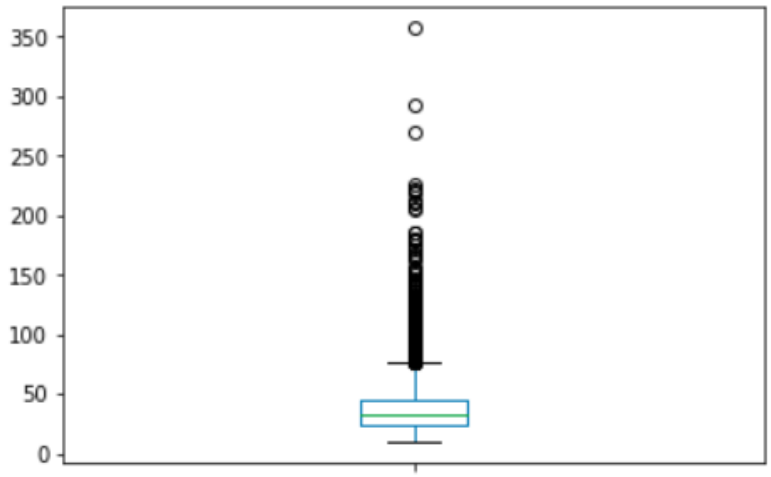


Figura 16 - Número de atividades por processo.

Além do número de atividades realizadas em cada um dos casos, outras informações devem ser levadas em consideração para entender quando o ticket será solucionado dentro do prazo esperado. As Tabelas 5, 6 e 7 mostram a relação de alguns dos atributos do chamado com seu prazo de resolução.

Tabela 5 - Descrição dos casos do log.

Total de tickets	Resolvidos dentro da SLA	Resolvidos fora da SLA
6337	3370	2967

Tabela 6 - SLA por prioridade.

Prioridade	Resolvidos dentro da SLA	Resolvidos fora da SLA
1	1	2
2	641	273
3	2214	2317
4	473	359
5	41	16

Tabela 7 - Chamados mais comuns.

Tipo do chamado	Resolvidos dentro da SLA	Resolvidos fora da SLA
Manutenção da estação de trabalho	741	527
Aplicativos	526	534
Desempenho da estação de trabalho	326	605
Problemas com impressão	207	163
Problemas com e-mail	92	133
Problemas de rede	110	65

Com base nessas três tabelas pode-se observar que, por mais que para todo log o número de casos positivos e negativos esteja balanceada, a mesma regra não se aplica quando algumas características são observadas separadamente. São os casos dos chamados com prioridade ‘2’ ou do tipo ‘Desempenho da estação de trabalho’, por exemplo. Sendo que no primeiro caso, mais de 2/3 dos tickets foram resolvidos dentro do prazo e no segundo, se observa quase a mesma relação, porém tendo mais casos negativos.

Na base de dados inicial, o sistema de gerenciamento insere diversas colunas representando os atributos associados a cada atividade. Elas contêm informações como a prioridade do ticket, tipo de serviço e e-mail, entre outros tipos de dados. Visando um melhor resultado e eficiência no tempo de processamento, somente colunas consideradas relevantes para o contexto foram mantidas. Essas colunas estão indicadas na Tabela 8, assim como o tipo de dado armazenado.

Tabela 8 - Tipos de dados do log.

Nome do atributo	Descrição
CaseID	Número de identificação do ticket no registro.
EventID	Número que identifica e representa a ordem em que a atividade foi executada durante o processo.
EventName	Classe que identifica e classifica os diferentes tipos de atividades possíveis no processo.
Priority_ID	Número usado para identificar a prioridade, de acordo com o impacto de cada ticket.
ServiceType	Classe que indica o tipo de problema para o atendimento.
a_body	Corpo dos e-mails trocados durante as atividades de cada ticket.
SLAMissed	Classe que indica para cada caso, se o problema foi resolvido dentro do prazo estabelecido, de acordo com as prioridades.

Colunas com a data de resolução e abertura dos incidentes não foram utilizadas, uma vez que a coluna “SLAMissed” indica se o caso foi resolvido ou não dentro do tempo esperado, não havendo necessidade de calcular o tempo total da carga de trabalho. Seguindo o objetivo desse projeto de verificar a combinação de dados estruturados e texto livre para a melhoria dos resultados de predição, somente o corpo do e-mail foi mantido

para a análise, uma vez que as informações de remetente e destinatário são consideradas dados semiestruturados.

Por possuir diversas informações não relevantes ao contexto do que é tratado durante a resolução dos tickets, uma limpeza no texto foi realizada. Apresentações e assinaturas, por exemplo, foram retiradas, assim como dados de telefone dos atendentes e diferentes links para o site da empresa. Além disso, um pré-processamento foi executado para remover as chamadas "palavras vazias".

Essas palavras são importantes para o entendimento do texto entre seus locutores, por serem utilizadas para definir gênero ou ligar as diferentes frases do texto, por exemplo, mas que para o algoritmo não acrescentam muito valor ao que se é aprendido, uma vez que são palavras muito comuns e, portanto, aparecem com muita frequência nas mensagens. Alguns exemplos são *'e'*, '*o'*, '*a'*, '*em'* e '*no'*. Além desse, outros tratamentos foram aplicados e serão detalhados na sequência.

A coluna utilizada como classe para previsão ('SLAMissed') tem seus valores positivos para indicar que o caso ultrapassou o prazo de resolução desejado e esses foram utilizado para apontar os casos de desvio, já que representam o oposto do que é esperado ao final da execução. Como citado anteriormente, a relação entre os dois casos é balanceada e, portanto, não foram necessárias etapas extras de balanceamento dos valores do log, que são utilizadas para evitar que o aprendizado de máquina seja tendencioso.

4.3. Análise Exploratória

Como o objetivo de aprimorar o processo de previsão, após o processamento do texto, uma nova coluna foi adicionada, contendo a quantidade de palavras utilizadas durante as mensagens para cada ticket. Dessa forma, é possível analisar também como o número de palavras presentes em cada caso afetam a resolução do problema. O gráfico na Figura 17 mostra a relação do número de palavras trocadas para os tickets que

obtiveram resultado positivo ao final da execução, enquanto o a Figura 18, para os desvios.

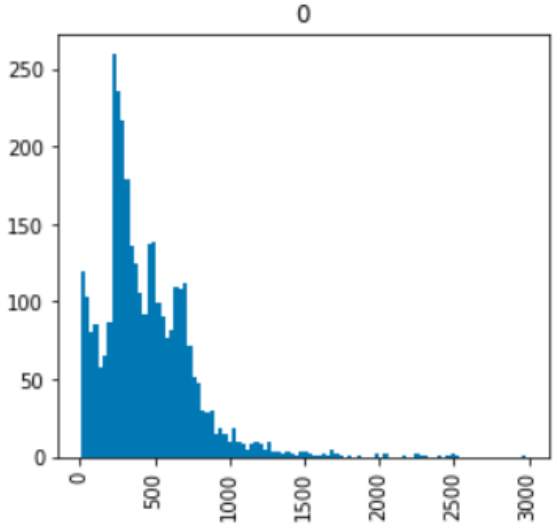


Figura 17 - Quantidade de mensagens por número de palavras.

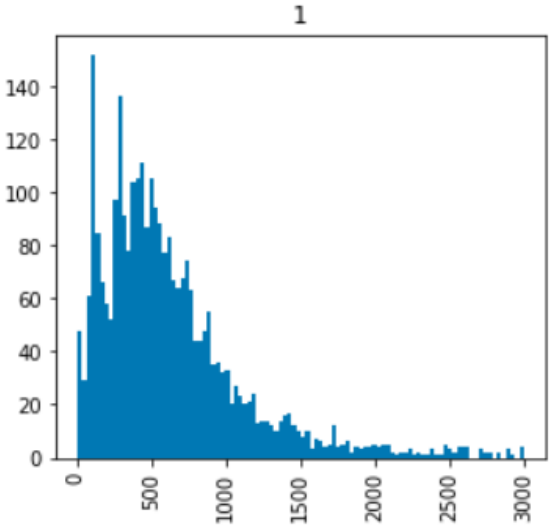


Figura 18 - Quantidade de mensagens por número de palavras.

No primeiro caso, observa-se que a grande maioria dos tickets possui até um valor próximo de 750 termos trocados e no segundo, esse valor se estende até um pouco mais do que 1000. A análise desses dois gráficos mostra uma leve relação entre a quantidade de palavras presentes e o seu tempo de resolução, uma vez que o caso de desvio apresenta mais tickets com número maior de vocábulos. Esse comportamento é reforçado na Figura

19, que mostra uma redução no número de casos dentro da SLA quando as palavras passam de 1900.

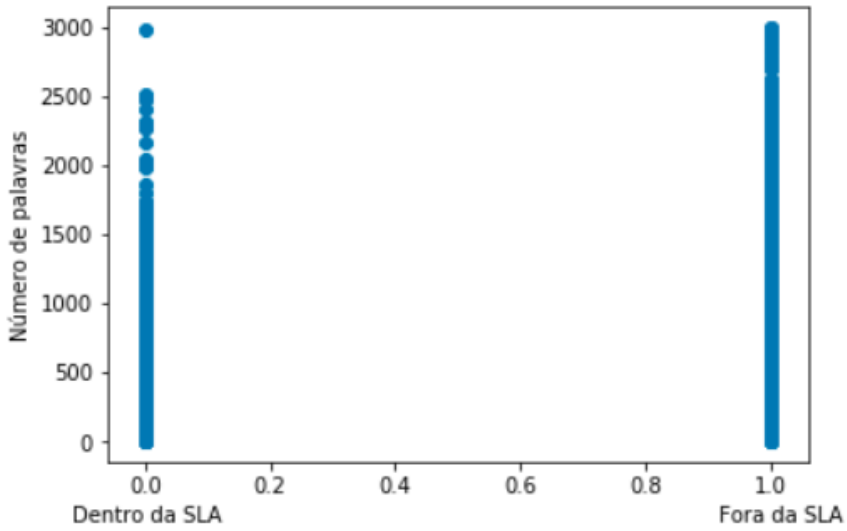


Figura 19 - Classificação por número de palavras.

Essas são algumas das análises diretas que podem ser feitas observando os dados relativos às mensagens de texto trocadas durante um atendimento. Elas exemplificam bem como a troca de informações na forma de linguagem natural pode afetar o resultado de um processo e reforçam a necessidade desse e outros trabalhos sobre o tema. Abaixo estão as etapas de execução, até o treinamento e teste dos princípios abordados.

4.3.1. Preparação do log

Após o pré-processamento do texto e adição de um novo atributo, mencionados anteriormente, a próxima etapa da limpeza foi a seleção dos atributos utilizados na predição. No log inicial existiam informações que não acrescentavam valor para o resultado visado no projeto, o de prever quais casos devem terminar ou não dentro da SLA estabelecida. Sendo assim, colunas com o início, o fim e a duração do atendimento foram retiradas, já que se obtém de maneira mais simples se cada caso teve seu prazo atendido ou não através da coluna “SLAMissed”. Seus valores foram usados como as classes previstas ao final da execução, sendo classificados casos positivos e negativos com os valores de Falso (ou 0) e Positivo (ou 1), respectivamente.

Seguindo essa ação, as categorias presentes em cada coluna foram transformadas em números, pois é dessa maneira que o algoritmo conseguirá processar as informações do log. As Figuras 20 e 21 mostram um exemplo de como o log fica antes e depois dessa etapa.

	serviceType	priority_id	a_body_group	SLAMissed
0	Redes::Diagnósticos	3	Prezado Paulo, bom dia!\r\n\r\n \r\n\r\nNão es...	N
62	Estações de Trabalho::Diagnósticos	3	Bom dia,\r\n\r\n \r\n\r\nNão estou conseguindo...	Y
94	Estações de Trabalho::Aplicativos	4	Cliente Jaqueline e Roberto informam que tem u...	N
121	Redes::Diagnósticos	3	Cliente Anderson informa de lentidão no acesso...	N
146	Estações de Trabalho::Impressão	2	Estação da Tamiris não imprime receitasEstação...	Y
166	Estações de Trabalho::Aplicativos	2	Problemas no caixaProblemas no caixa\r\n \r\n...	Y
186	Estações de Trabalho::Manutenção	3	\r\nBom dia !!!!\r\n\r\nComputador da vendedor...	Y
250	Estações de Trabalho::Manutenção	2	Cliente informa que o engenheiro Jorge Roberto...	N
275	Redes::Diagnósticos	4	Cliente informa que os usuários estão sem aces...	N
321	Correio eletrônico::Manutenção	3	Verificar email da Roberta Kfuri, que desconfi...	N

Figura 20 - Log com categorias em texto.

	serviceType	priority_id	a_body_group	SLAMissed
0	55	3	Prezado Paulo, bom dia!\r\n\r\n \r\n\r\nNão es...	0
62	21	3	Bom dia,\r\n\r\n \r\n\r\nNão estou conseguindo...	1
94	19	4	Cliente Jaqueline e Roberto informam que tem u...	0
121	55	3	Cliente Anderson informa de lentidão no acesso...	0
146	23	2	Estação da Tamiris não imprime receitasEstação...	1
166	19	2	Problemas no caixaProblemas no caixa\r\n \r\n...	1
186	24	3	\r\nBom dia !!!!\r\n\r\nComputador da vendedor...	1
250	24	2	Cliente informa que o engenheiro Jorge Roberto...	0
275	55	4	Cliente informa que os usuários estão sem aces...	0
321	13	3	Verificar email da Roberta Kfuri, que desconfi...	0

Figura 21 - Log com categorias enumeradas.

Após a preparação inicial do log, os próximos passos se deram através do processamento do texto e criação do vetor de características. Durante o desenvolvimento do projeto o conteúdo do texto foi utilizado para extração desses vetores e isso quer dizer que o tamanho desse mesmo vetor será dado de acordo com o tamanho do dicionário de palavras selecionado. Dessa forma, por se tratar de documentos que podem conter um grande número de palavras, as características extraídas irão apresentar um grande volume,

criando vetores muito esparsos, uma vez que diversas palavras devem aparecer poucas vezes em textos diferentes. Isso reduz o desempenho, pois o algoritmo terá que percorrer cada um e irá aprender pouco com os vários valores nulos que estarão representados.

Visando melhorar o desempenho de processamento e trazer informações relevantes ao conteúdo, uma limpeza e seleção de palavras foi realizada. Ela seu deu da seguinte forma:

1- Extração do texto não relevante para a resolução dos incidentes: Normalmente, e-mails apresentam uma estrutura similar para cada tipo de usuário e empresa, como assinaturas, sites web, telefones de contato e apresentações. Nessa etapa, o objetivo foi retirar esse tipo de informação, que normalmente não fazem parte do contexto descritivo do problema e nem das atividades realizadas durante o processo. Além de excluir informações que não agregam para a previsão, a retiradas desses termos também auxilia no desempenho do algoritmo, pois o número de palavras lidas será menor. No exemplo (1) pode-se ver um corpo do e-mail como seria registrado no log:

E na sequência, como o mesmo texto ficaria somente após o primeiro filtro das (1) informações.

Não consigo enviar mensagens do meu e-mail corporativo.
Não está funcionando para receber novas mensagens também.

2- Retirada de pontuação: Para que seja possível passar a entonação e consequentemente, dar o sentido desejado ao que está escrito, a pontuação se faz presente nos e-mails trocados. Porém, para o modelo de texto a ser criado, esses caracteres não agregam valor, além de aumentarem o número de informação a ser

computada. Nesse segundo passo, a tarefa foi retirar esses marcadores antes de separarmos as palavras dos textos.

Bom dia!

Não consigo enviar mensagens do meu e-mail corporativo.
Não está funcionando para receber novas mensagens também.

Podem verificar, por favor?

Att,
Nome Sobrenome
Analista de Marketing
Tel.:(xx) 0000-0000

3- Retirada das 'Palavras Vazias': Mais uma vez, para dar sentido a escrita, diversas palavras são utilizadas, como os conectores ou artigos para definição do gênero e consequentemente não refletem o conteúdo da conversa. Durante essa etapa esses vocábulos foram retirados. A mensagem passaria a ser representada como abaixo:

Não consigo enviar mensagens e-mail corporativo
Não funcionando receber novas mensagens

4- Criação de *tokens*: Etapa chave para a criar o modelo de texto, é nesse momento que as palavras são separadas e adicionadas às posições no vetor. Uma vez que houve a retirada de pontuação, as palavras estão separadas através dos espaços em branco no texto. Para que o algoritmo não computasse separadamente as palavras somente por conter letras maiúsculas, todas foram normalizadas para minúsculo. No exemplo, um vetor x, contendo os *tokens* retirados do texto teria seus valores indicados por x = {'não', 'consigo', ..., 'novas', 'mensagens'}.

5- *Stemming*: O termo em inglês é utilizado em processamento de linguagem natural, e define a prática de se extrair o radical dos termos presentes no texto. Em outras palavras, as diferentes formas das palavras derivados da mesma raiz, são todas reduzidas a esta. Para exemplificar, os vocábulos "apresentar", "apresentação", "apresentando", seriam reduzidas à "apresent", uma vez que somente os sufixos são diferentes. Como cada língua apresenta características específicas, a função stemm do Python para português foi utilizada. Seguindo a nova etapa, as

informações no vetor ficariam x = {“não”, “conseg”, ..., “funciona”, “receb”, “nova”, “mensage”}

Esse tratamento das informações contidas nas mensagens de texto foi importante para evitar conteúdo que interferisse no resultado negativamente. Além disso, ajudou a reduzir a quantidade de informação presente nos documentos de texto, melhorando a velocidade de processamento. O uso desses dados para extração do vetor de características e treinamento dos classificadores estão detalhados a seguir

4.3.2. Extração dos vetores de características

Uma vez que cada uma das mensagens está representada por uma lista de palavras, ou ‘*tokens*’, o próximo passo foi convertê-las em um vetor de características de texto. No contexto de extração de informação e aprendizado de máquina, uma característica é definida como uma propriedade mensurável individual ou uma característica de um fenômeno sendo observado e é utilizada para reconhecimento de padrões. Como os métodos de aprendizado de máquina não trabalham diretamente com o texto bruto, o vetor de características de texto foi usado como valor de entrada desses métodos. Como a redução do número de palavras presentes nos documentos de texto apresentou uma perda considerável no desempenho da predição e não teve ganhos significativos no tempo de processamento do algoritmo, o dicionário de palavras criado conteve as mesmas palavras presentes nos documentos de texto.

Seguindo as etapas do modelo *bag of words*, foram contadas quantas vezes cada uma das palavras aparecem em cada documento. Isso quer dizer que cada uma delas corresponde à uma coluna da tabela e os valores em cada linha representam a quantidade de vezes que elas aparecem no documento associado a cada atividade. Sendo assim, atividades que não tiveram troca de mensagem terão seus valores iguais a zero. Um exemplo para um vetor com as seguintes palavras {“não”, “receb”, “mensage”, “alerta”} poderia ser representado como na Tabela 9.

Tabela 9 - Log com frequência de palavras.

Ticket ID	Prioridade	Frequência das palavras				
		não	Receb	mensage	alerta	Dentro da SLA
256	3	0	1	1	1	S
257	2	1	1	1	0	N
365	3	2	2	1	1	N

Com base na tabela, pode-se ver a quantidade de vezes que cada uma das palavras aparece em cada ticket. No primeiro caso, por exemplo, todos os termos aparecem somente uma vez, com exceção do “não”.

4.3.3. Aplicação do *Transformer TF-IFD*

Após criar o vetor de características, o próximo passo foi normalizar os valores encontrados para cada termo, pois, como explicado anteriormente, existem palavras que podem aparecer diversas vezes nos documentos, mas que não adicionam informações relevantes para o contexto. Caso não sejam tratadas, essas palavras podem sobrepujar outras que não aparecem tantas vezes, mas que podem ser especificas sobre o domínio estudado, por exemplo.

Para que esses dados não dominassem o modelo com uma alta contagem e influenciassem a previsão de forma negativa, a técnica de pontuação TF-IDF foi utilizada. Essa abordagem compara a frequência com que as palavras aparecem nos documentos de texto separadamente com o quão raras essas palavras são entre todos os outros documentos de texto existentes. Nesse caso, as expressões frequentes no corpo do e-mail analisado, e em todos os outros documentos, têm sua pontuação penalizada. O efeito disso é que palavras distintas receberão um peso maior, se comparadas àquelas que são muito utilizadas. A partir de então, os termos deixam de ser representados pela sua frequência no documento e passam a ser representados pelo seu valor ponderado.

Após essas etapas, o log está pronto para o aprendizado, já que possui seus valores representados por vetores de números de tamanho fixo. A combinação dos valores para

as variáveis estruturadas e do texto processado de cada linha foi utilizada pelo modelo como treino e teste para prever o valor da classe associada como resultado. Em outras palavras, o valor de cada linha da coluna “SLAmissed” irá apresentar uma combinação de eventos, de atributos associados e de palavras presentes no corpo dos e-mails, que será utilizada para mapear o comportamento padrão através dos algoritmos de aprendizado de máquina.

4.3.4. Treinamento dos classificadores

Para treinar os classificadores, após toda a etapa de preparação, os registros do log foram separados em grupos diferentes, para que o conjunto de dados de treinamento fosse diferente do utilizado para teste e não fosse enviesado. Essa separação dos registros do log foi feita utilizando-se 75% dos valores existentes para o conjunto de treino e 25% para o conjunto de teste.

Para criar um baseline para comparação dos resultados do projeto, foram utilizados os valores obtidos através do treinamento e teste utilizando-se somente os dados estruturados do processo. Dessa forma, é possível comparar se a combinação dos diferentes atributos realmente apresenta uma melhora de resultado e validar as hipóteses levantadas durante a fase inicial. Para comparação, outro teste foi realizado considerando somente as características extraídas do corpo dos e-mails.

4.3.4.1. Floresta aleatória

Assim como indicado no paragrafo anterior, o modelo que utilizou o algoritmo de florestas aleatórias também foi inicialmente treinado utilizando-se somente as colunas que continham dados estruturados. Os resultados obtidos indicam que o algoritmo não apresenta um bom desempenho para a predição nessa configuração, pois o F-score apresentado é muito próximo de 50%. Ou seja, isso indica que usar essa abordagem será quase tão assertiva quanto um palite aleatório. A Tabela 10 representa a matriz de confusão para o caso.

Tabela 10 - Matriz de confusão – dados estruturados (FA).

Real / Previsto	Dentro da SLA	Fora da SLA
Dentro da SLA	528	313
Fora da SLA	328	416

Nas Tabelas 11 e 12, são apresentados os resultados para o caso em que os dois tipos de dados, não estruturados e estruturados, foram combinados para realizar a previsão.

Tabela 11 - Matriz de confusão – dados estruturados e não estruturados (FA).

Real / Previsto	Dentro da SLA	Fora da SLA
Dentro da SLA	640	201
Fora da SLA	293	451

Tabela 12 - Relatório de classificação - dados estruturados e não estruturados (FA).

	Precisão	Recall	F-score
Dentro da SLA	0,69	0,79	0,74
Fora da SLA	0,71	0,61	0,66
Media/Total	0,70	0,70	0,70

Esses resultados apontam uma melhora em relação ao primeiro caso. Na matriz de confusão, por exemplo, para ambos os casos reais, a quantidade de valores previstos corretamente, aumentou. Esse comportamento é mais acentuado para os casos positivos, quando o modelo foi treinado somente com dados estruturados, acertou 528 do total de 841 e no segundo caso, esse número passou a ser de 640, o que explica os valores encontrados para o recall em ambos.

Um outro comportamento observado é o de que o algoritmo utilizado possui um melhor desempenho para analisar os casos que não apresentam desvio, ou seja, esse

método será mais preciso quando o objetivo for monitorar os tickets que serão finalizados dentro da SLA. Como no contexto em que o projeto está inserido o objetivo é o de monitorar e tentar prever os casos que irão ultrapassar o prazo inicial (para auxílio a tomada de decisão), esse resultado indica que esse pode não ser o melhor algoritmo, mesmo que a diferença entre os casos não seja tão grande.

4.3.4.2. Regressão logística

Assim como o que aconteceu com o algoritmo de floresta randômica, os resultados dos modelos aprendidos com base na combinação de ambos os dados, apresentou um melhor desempenho quando comparado à *Baseline*. As Tabelas 13 e 14 apresentam os valores obtidos utilizando dados estruturados.

Tabela 13 - Matriz de confusão – dados estruturados (RL).

Real / Previsto	Dentro da SLA	Fora da SLA
Dentro da SLA	729	112
Fora da SLA	649	95

Tabela 14 - Relatório de classificação – dados estruturados (RL).

	Precisão	Recall	F-score
Dentro da SLA	0,53	0,87	0,66
Fora da SLA	0,46	0,13	0,20
Media/Total	0,50	0,52	0,44

A *Baseline* para esse algoritmo apresenta um resultado muito diferente do foi obtido com o primeiro algoritmo. Se vê uma grande melhora em relação a previsão dos casos reais positivos Recall de 0,87 (comparado aos 0,67 do primeiro algoritmo), mas que ao mesmo tempo apresenta um piora acentuada em relação aos casos falso positivos com bem abaixo do primeiro caso, 0,13. Ou seja, para essa configuração, a regressão logística tende a classificar os casos como não desviantes, o que, na média entre os valores para

casos positivos e negativos, faz com que seu resultado acerte em menos do que 50% dos casos (F-score = 0,44), sendo pior do que um palpite aleatório.

Como indicado no início do tópico, o resultado da combinação dos atributos também apresenta progresso. Como a *Baseline* desse algoritmo teve desempenho muito baixo, a melhora é ainda mais significativa. As Tabelas 15 e 16 detalham esses valores.

Tabela 15 - Matriz de confusão – dados estruturados e não estruturados (RL).

Real / Previsto	Dentro da SLA	Fora da SLA
Dentro da SLA	663	178
Fora da SLA	281	463

Tabela 16 - Relatório de classificação - dados estruturados e não estruturados (RL).

	Precisão	Recall	F-score
Dentro da SLA	0,70	0,79	0,74
Fora da SLA	0,72	0,62	0,67
Media/Total	0,71	0,71	0,71

Observando os valores da matriz de confusão, pode-se ver que o número de casos corretamente previstos como positivos não sofre grande alteração, 729 no primeiro caso e 663, no segundo, porém existe uma grande mudança em relação aos casos de desvio. Nessa nova configuração, a quantidade de falso positivos diminuiu drasticamente, passando de 649 do total de 744 casos de desvio, para 281. Dessa forma, o resultado melhora consideravelmente, tendo F-score melhor até do que o que se obtém com floresta randômica.

Assim como o primeiro algoritmo, a regressão logística consegue prever com mais precisão os casos não desviantes, porém com uma diferença menor em relação a esse para os tickets cuja resolução que fogem da SLA pré-estabelecida. Isso mostra que, apesar de

ainda não ser ideal, funciona melhor para previsão em um log com as características do contexto desse projeto.

4.4. Discussão

Como discutido em outros capítulos, algumas características influenciam diretamente no tempo de resolução dos chamados. Um exemplo disso é a quantidade de atividades necessárias até que se chegue a resolução do problema. Além dessa característica presente no processo, as informações atreladas a essas atividades também podem exercer grande influência, como tipo do ticket ou a fila de atendimento que ele será associado, visto que pessoas diferentes operam de modo diferente.

Nesse projeto o objetivo foi verificar como as mensagens trocadas durante a execução de cada ticket influenciam no tempo de resolução e entender como alguns dos algoritmos de aprendizado de máquina se comportam para um log com características semelhantes. No contexto descrito, a base de dados apresenta uma quantidade balanceada de valores positivos e negativos para o que se deseja prever.

Outro ponto é que, diferentemente do apresentado por Teineema et al. (2016), o corpo dos e-mails associados às atividades é escrito pelos próprios usuários e não seguem nenhuma estrutura padrão. Assim, foi possível estudar as hipóteses levantadas durante o artigo citado para o comportamento em textos mais heterogêneos.

Uma das questões é que o artigo de Teinemaa informa que os resultados obtidos com floresta randômica foram melhores quando comparados aos da regressão logística, o que não foi observado durante esse projeto. Por mais que a *Baseline* tenha apresentado um resulto ruim para o algoritmo, sua previsão apresenta resultado tão bom quanto o de floresta randômica, ao combiná-los com as informações extraídas dos textos.

Com base nas análises feitas durante a apresentação das matrizes de confusão e relatórios de classificação, observa-se que, mesmo com comportamento intermediário diferente para cada um dos casos, o resultado reforça que a utilização de dados extraídos de texto ajudam na eficiência para previsão do valor de uma classe em um determinado

momento da execução do processo em andamento, já que em todos os casos o valor para F-score apresentou grande melhora.

5. Conclusão

Após a apresentação dos resultados e de sua análise, esse capítulo irá resumir a comparação do resultado esperado ao obtido, apontar as principais dificuldades encontradas e os potenciais temas de sequência e melhoria do que foi estudado durante esse projeto.

5.2. Considerações finais

Como descrito no primeiro capítulo deste documento, o principal objetivo do projeto foi o de reforçar ou refutar a hipótese de que a combinação de características extraídas de texto livre e atributos estruturados inerentes ao processo melhora a eficiência na predição de resultado para os casos em andamento. Para alcançar tal objetivo, outros estudos foram utilizados como base da pesquisa, em especial um artigo que realizou testes com diversos modelos de texto para processamento de linguagem natural e algoritmos de aprendizagem de máquina. Os resultados e análise desse artigo serviram como ponto de partida para as definições do projeto. Um exemplo disso foi a escolha do modelo de texto a ser utilizado.

Por mais que o comportamento dos algoritmos tenha sido uma pouco diferente do mencionado pela referência, os resultados aqui obtidos reforçam a hipótese de que essa combinação melhora consideravelmente a predição em processos de negócio. A manipulação de dados não-estruturados assume um papel cada vez mais importante, onde a quantidade e velocidade de mensagens trocadas na internet aumenta gradativamente.

5.3. Limitações

Um dos principais pontos de dificuldade ao se trabalhar com texto livre, é a quantidade de informação extra existente nos documentos estudados e para isso existem diversas técnicas de limpeza, com o objetivo que o mínimo possível do conteúdo apresente dados que não são relevantes para o processo.

Neste projeto, diversas dessas técnicas foram aplicadas para eliminar as palavras irrelevantes ao contexto. Porém, por ser tratar da troca de e-mails entre o suporte e o

cliente, muitas das mensagens analisadas possuíam parte da sequência anterior, dentro do mesmo corpo. Caso isso fosse aplicado para todos os casos, uma solução seria a seleção somente do último e-mail trocado. Como isso não se aplica ao log, todas as mensagens foram mantidas para que não houvesse a perda de conteúdo, porém essa solução gera tópicos repetidos.

5.4. Trabalhos futuros

Para sequência ao que foi estudado durante este projeto, alguns tópicos são sugeridos:

- Comportamento de outros algoritmos de aprendizado máquina e de outros modelos de texto em um log com características semelhantes para analisar quais algoritmos e modelos são mais indicados para predição no mesmo tipo de contexto;

- Análise do impacto de descrições erradas dos problemas para a resolução dos chamados para verificar se é possível encontrar um padrão relacionado a isso no texto e aprimorar os resultados de predição, retirando-se esse tipo de informação do texto.

Referências Bibliográficas

Aalst, W. M. P. (2011) Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer.

Aalst, W. M. P. (2012). Process mining: Overview and opportunities. ACM Trans. Manage. Inf. Syst. 3, 2, Article 7 (July 2012).

Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L. (2004). "Workflow Mining: Discovering Process Models from Event Logs", IEEE Transactions on Knowledge and Data Engineering, 16(9):1128–1142.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. and Kochut. K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. In Proceedings of KDD Bigdas, Halifax, Canada, August 2017, 13 pages.

Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of machine Learning research 3 (2003), 993–1022.

Breiman, L. (2001), Random forests. Machine learning 45(1), 5-32.

Carvalho, D; Moser, A; Silva, V, Dallagassa, M. Mineração de Dados aplicada à fisioterapia. Fisioter. mov. [online]. 2012, vol.25, n.3 [cited 2018-10-31], pp.595-605.

Castro, L. Introdução à Mineração de Dados. Material de apoio do curso Mineração de Dados do PPGEE-Universidade Mackenzie, disponível online em https://pt.slideshare.net/lndecastro/, pag 403.

Conforti R., Fink S., Manderscheid J., Röglinger M. (2016). PRISM – A Predictive Risk Monitoring Approach for Business Processes. In: La Rosa M., Loos P., Pastor O. (eds) Business Process Management. BPM 2016. Lecture Notes in Computer Science, vol 9850. Springer, Cham.

Conforti, R., de Leoni, M., Rosa, M. L., Aalst, W. M. P. (2013). Supporting riskinformed decisions during business process execution. In: Proc. of CAiSE. pp. 116–132.

Dumas, M., La Rosa, M., Mendling, J., Reijers, H.A. (2013). Fundamentals of Business Process Management. Springer.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. DOI: https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U; Piatetsky-Shapiro, G; Smyth, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.

Feldman, R. and Sanger, J. (2002), The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 1st edition.

Francisco, R. and Portela Santos, E. A. (2011). Aplicação da Mineração de Processos como uma prática para a Gestão do Conhecimento. Simpósio Brasileiro de Sistemas de Informação (SBSI). Salvador, BA. 447-484.

Freedman, D. (2005) Statistical Models: Theory and Practice. Cambridge University Press.

Jagannath, V. Random Forest Template for TIBCO Spotfire® – Wiki page. Disponível em https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page. 2017

Maggil, F.M., Francescomarino, C. D., Dumas, M., Ghidini, C. (2013). Predictive Monitoring of Business Processes. arXiv:1312.4874v2 [cs.SE], 19 Dec 2013.

Marquez-Chamorro, A. and Resinas, M. and Ruiz-Cortés, A. (2017). Predictive monitoring of business processes: a survey. IEEE Transactions on Services Computing. PP. 1-1. 10.1109/TSC.2017.2772256.

Oliveira, M. e Bertucci, M. G. E. S. (2003). A pequena e média empresa e a gestão da informação. Informação & Sociedade: Estudos, João Pessoa, v. 13, n. 2.

Pesic, M., Aalst, W.M.P. (2006). A Declarative Approach for Flexible Business Processes Management. In: BPM Conference 2006 Workshops. pp. 169–180.

Quilan, J.R. (1986). Induction of Decision Trees. Machine Learning 1: 81-106, 1986 Kluwer Academic Publishers, Boston

Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review. 33 (1-2): 1–39, 2010.

Senderovich, A., Shleyfman, A., Weidlich, M., Gal, A. and Mandelbaum, A. (2016). P3-Folder: Optimal Model Simplification for Improving Accuracy in Process Performance Prediction. BPM 2016, LNCS 9850, pp. 418-436.

Tan, A.H. (1999). Text Mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, 1999.

Teinemaa, I.; Dumas, M.; Maggil, F. M.; Francescomarino, C. D. (2016). Predictive Business Process Monitoring with Structured and Unstructured Data. In Business Process Management, pages 401–417.

Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54: 167–178. doi:10.2307/2333860.

Weiss, S. I. and Kulikowski, C. (1991). Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, California: Morgan Kaufmann.