# Mineração de Padrões Sequenciais de Sentimentos: Um Estudo de Caso na Prevenção de Evasão da Educação Superior

**Thiago Pimentel**[1], **Claudio Passos**[1], **Isabel Fernandes**[2], **Ronaldo Goldschmidt**[1]

[1]Seção de Engenharia da Computação (SE/9) – Instituto Militar de Engenharia (IME)
22.290-270 – Rio de Janeiro – RJ – Brasil

[2]Campus Foz do Iguaçu - Instituto Federal do Paraná (IFPR)
85.860-000 - Foz do Iguaçu - PR - Brasil

*pimentel@ime.eb.br, cpassos.cp2@gmail.com, ifsouza@yahoo.com.br, ronaldo.rgold@ime.eb.br*

***Abstract.*** *In view of the high dropout rates of Brazilian undergraduate courses, this paper investigates the hypothesis that the use of the underlying sentiment in the student-university interactions can improve sequential pattern mining based dropout prediction. To this end, the present work applied an adapted version of the method proposed in SS-DetChurn to a set of historical data of a university. Quantitative results of the experiments confirmed the hypothesis raised. In addition, the patterns discovered led to a set of actions that can be added to the university's dropout combat program.*

***Resumo.*** *Em um cenário de altos índices de evasão na Educação Superior brasileira, o presente trabalho investiga a validade da hipótese de que utilizar informações sobre possíveis sentimentos presentes nas interações entre aluno e instituição de ensino superior (IES) pode contribuir para melhorar a detecção antecipada de evasão escolar baseada na mineração de padrões sequenciais. Para tanto, adaptou o método proposto em SS-DetChurn de forma a aplicá-lo sobre dados históricos de uma IES. Os experimentos realizados com esses dados produziram resultados quantitativos que apontaram para a validade da hipótese investigada e resultados qualitativos que levaram a um conjunto de ações a ser incorporado no programa de combate à evasão da referida IES.*

## 1. Introdução

A evasão escolar na Educação Superior brasileira vem crescendo significativamente nos últimos anos. Segundo levantamento divulgado pelo SEMESP[1] em 2018, a evasão em cursos de graduação presencial no Brasil atingiu 30,1% na rede privada e 18,5% na rede pública [Bocchini 2018]. Nos cursos de educação a distância (EaD), o índice chegou a 36,6% na rede privada e a 30,4% na pública [Bocchini 2018]. Diante de estatísticas tão expressivas, ganham relevância as iniciativas de pesquisa voltadas ao desenvolvimento e aplicação de instrumentos computacionais que possam ser utilizados para mitigar a ocorrência de evasão escolar no referido nível de ensino.

A fim de se relacionar com seus alunos de forma mais próxima e personalizada, e, consequentemente procurar reduzir a evasão escolar, diversas instituições de ensino

[1]Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo.

DOI: 10.5753/cbie.sbie.2019.1411

brasileiras têm investido na implantação de ambientes computacionais de *CRM* (*Client Relationship Managemenent*) - usado pelas organizações para criar, desenvolver e aprimorar os relacionamentos com clientes. Em geral, o *CRM* abrange canais de comunicação tais como *e-mail*, *chat*, telefone, dentre outros, que armazenam registros sobre todas as conversas ocorridas ao longo do tempo [Antonucci 2018].

Um dos principais desafios enfrentados pelo *CRM* das organizações, independente do segmento em que atuam, é a identificação de clientes com propensão ao *churn* (i.e., cancelamento de produtos e/ou serviços) [Hadden et al. 2007]. No *CRM* das instituições do segmento educacional, o *churn* pode ser interpretado como evasão de aluno. Consequentemente, o problema de se identificar alunos com propensão à evasão pode, por sua vez, ser interpretado como o problema de se detectar antecipadamente a ocorrência de *churn*. Assim, uma vez identificados os alunos com maior propensão ao *churn*, as instituições podem intervir de forma ágil junto a esses alunos a fim de evitar sua evasão.

Diversos trabalhos de pesquisa têm buscado a criação de modelos que tentam detectar antecipadamente a ocorrência de *churn* [García et al. 2017], independente do segmento onde o *CRM* esteja inserido. A maioria desses trabalhos envolve a aplicação de técnicas de aprendizado de máquina sobre dados de perfil dos clientes e de suas interações com as organizações. Algumas dessas pesquisas têm obtido melhor desempenho que as demais por utilizarem métodos de detecção de padrões sequenciais de comportamento [Chiang et al. 2003]. Tais métodos consideram a ordem cronológica e os tipos de interações entre o cliente e a empresa que precedem os eventos de *churn*, para identificar padrões de propensão ao desligamento. Embora promissores, tais métodos deixam de considerar um aspecto muitas vezes presente nas conversas registradas entre cliente e empresa e que pode fornecer indícios importantes para criação de modelos de prevenção de *churn*: os sentimentos manifestados durante as interações ocorridas ao longo do tempo. Diante deste cenário, em [Pimentel and Goldschmidt 2019], os autores propuseram o *SS-DetChurn*, um método de detecção de padrões sequenciais que considera os sentimentos extraídos das interações entre clientes e organizações. Segundo o referido artigo, os experimentos realizados indicaram a adequação do *SS-DetChurn* quando aplicado em dados do *CRM* de empresas do segmento de telecomunicações.

Diante do exposto, o presente trabalho levanta a hipótese de que utilizar informações sobre possíveis sentimentos presentes nas interações entre aluno e instituição de ensino na ordem em que elas ocorrem ao longo do tempo pode melhorar a prevenção de *churn* baseada na detecção de padrões sequenciais em ambientes educacionais. Assim, este artigo tem como objetivo apresentar evidências experimentais que confirmem a hipótese levantada. Para tanto, o artigo reporta as adaptações necessárias ao *SS-DetChurn* de forma a aplicá-lo sobre informações acerca das interações ocorridas longitudinalmente entre alunos e instituição de ensino. Nos experimentos com dados históricos sobre os contatos realizados junto aos alunos de uma instituição de ensino superior, o *SS-DetChurn* produziu resultados quantitativos que apontam para a validade da hipótese formulada. Adicionalmente, a partir dos dados qualitativos obtidos nos mesmos experimentos, o artigo propõe e discute ainda um conjunto de ações a ser incorporado no programa de combate à evasão escolar da referida instituição.

O presente texto possui mais cinco seções. A Seção 2 apresenta fundamentos sobre detecção de padrões sequenciais, necessários à compreensão do método adotado no

estudo de caso da pesquisa. Na Seção 3, são apresentados os trabalhos relacionados ao tema. A descrição formal da versão adaptada do *SS-DetChurn* encontra-se na Seção 4. Detalhes sobre os experimentos realizados e as análises sobre os resultados quantitativos e qualitativos obtidos estão na Seção 5. Por fim, na Seção 6, são destacadas as principais contribuições deste trabalho e indicados possíveis trabalhos futuros.

## 2. Mineração de regras de associação e de padrões sequenciais

A Mineração de Regras de Associação consiste em identificar regras de associação frequentes e válidas em um conjunto de dados [Agrawal et al. 1993]. Uma regra de associação *R* é uma implicação da forma $X \rightarrow Y$, onde *X* e *Y* são conjuntos de itens tais que $X \cap Y = \emptyset$. Um item é uma condição que pode assumir valor verdadeiro ou falso em função do registro de dados selecionado. Por exemplo, $R_1 : \{Turno = Integral, PraticaEsporte = Sim\} \rightarrow \{DesempenhoEscolar = Bom\}$ é uma regra de associação onde $Turno = Integral$, $PraticaEsporte = Sim$ e $DesempenhoEscolar = Bom$ são itens. Satisfazem a $R_1$, todos os registros do conjunto de dados que satisfazem (i.e., tornam verdadeiros) esses três itens ao mesmo tempo. Uma regra de associação $R : X \rightarrow Y$ é dita frequente (resp. válida) se, e somente se, $Sup(R) = |X \cup Y|/|D| \geq$ ***MinSup*** (resp. $Conf(R) = |X \cup Y|/|X| \geq$ ***MinConf***), onde $X \cup Y$ é o conjunto de registros que satisfazem aos itens em *X* e aos itens *Y* simultaneamente; $Sup(R)$ e $Conf(R)$ são, respectivamente, o suporte e a confiança de $R$; $|D|$ representa a quantidade total de registros de dados disponíveis no conjunto de dados $D$; e ***MinSup*** (suporte mínimo) e ***MinConf*** (confiança mínima) são parâmetros, de escolha no conjunto de treino, definidos pelo usuário. O conjunto foi utilizado nos experimentos do estudo de caso.

Inspirada nos conceitos acima, a Mineração de Padrões Sequenciais é uma extensão da tarefa de Mineração de Regras de Associação a fim de identificar sequências de eventos frequentes ocorridos ao longo do tempo. A seguir encontra-se uma descrição resumida das definições extraídas de [Pimentel et al. 2019] e [Pimentel and Goldschmidt 2019], que são a base para a busca de padrões sequenciais frequentes. A primeira delas é que cada registro de dados $r$ corresponde à ocorrência de um evento e é caracterizado por uma tripla ordenada da forma $r = (t, o, s)$, onde $t$ indica o momento de ocorrência do evento, $o$ é o elemento responsável pela ocorrência de $r$ e $s$ é um conjunto de itens que descrevem o evento $r$. No contexto do presente trabalho, cada registro de dados $r$ retrata um evento de interação (i.e., contato) entre um aluno $o$ e sua instituição de ensino ocorrido em um momento $t$. Informações sobre a interação registrada em $r$ são representadas em $s$. Assim sendo, define-se sequência de eventos associados a um elemento $o$ como um conjunto ordenado de registros de dados da forma: $S_o = \{r_i = (t_i, o_i, s_i) \mid o_i = o,\ i \in \{1, 2, ..., n\}\}$, ou, simplificadamente, $S_o = \ll s_{k_1}, s_{k_2}, ..., s_{k_r} \gg$, onde $k_1 < k_2 < ... < k_r$. A ideia básica desta tarefa de mineração é identificar sequências $P$ frequentes e válidas. Para tanto, dada uma sequência $P = \ll p_1, p_2, ..., p_r \gg$ (também denominada padrão sequencial), define-se suporte de $P$ ($Sup(P)$) como sendo o número de objetos aos quais $P$ é uma sequência associada. $P$ é considerada um padrão sequencial frequente se, e somente se, $Sup(P) \geq MinSup$. Adicionalmente, define-se confiança de $P$ como $Conf(P) = Sup(\ll p_1, p_2, ..., p_{r-1} \gg)/Sup(\ll p_r \gg)$. Assim, diz-se que $P$ é um padrão sequencial válido se, e somente se, $Conf(P) \geq MinConf$. ***MinSup*** e ***MinConf***

são parâmetros definidos pelo usuário.

## 3. Trabalhos relacionados

Conforme pode ser observado na Tabela 1, diversas iniciativas de pesquisa têm buscado a criação de modelos de predição de evasão escolar. Em geral, elas utilizam algoritmos de aprendizado de máquina para construir modelos de classificação que, diante de informações sobre perfil do aluno (e.g., idade, disciplina cursada, notas, frequência, dentre outras) buscam inferir se o aluno deverá ou não abandonar os estudos. Diferentemente da abordagem proposta no presente trabalho, nenhuma das iniciativas apresentadas interpreta o problema de predição de evasão escolar como um problema de detecção antecipada de *churn*. Mais do que isso, elas não levam em consideração importantes indícios para criação de modelos de predição de evasão: os sentimentos manifestados pelos alunos durante suas interações com as instituições e nem o aspecto temporal quanto à ocorrência dessas interações. Desta forma, deixam de considerar, por exemplo, que manifestações sucessivas e recorrentes de insatisfação de um aluno com aspectos de seu contexto educacional podem ser valiosos indicativos de tendência à evasão desse aluno.

**Tabela 1. Visão Resumida dos Trabalhos Relacionados**

| Referências | Padrões Sequenciais | Análise de Sentimentos | Algoritmos | Aplicação em Educação |
|---|---|---|---|---|
| [Manhães et al. 2012] | não | não | Naive Bayes | sim |
| [Junior et al. 2017] | não | não | Classificação, Criação e Seleção de atributos | sim |
| [Medeiros and Padilha 2018] | não | não | Algoritmos de classificação Part, OneR, J48 e Randomtree | sim |
| [Bezerra et al. 2016] | não | não | Árvore de Decisão, Indução de Regras e Regressão Logística | sim |
| [Gonçalvez et al. 2018] | não | não | Naive Bayes, Support Vector Machine e J48 | sim |
| [Colpani 2018] | não | não | Análise de Correlação e Regressão Linear | sim |
| [Cordeiro 2017] | não | não | J48 | sim |
| [Ramos et al. 2018] | não | não | Árvore de Decisão, Máquina de Vetor de Suporte, Rede Neural Artificial, k-Nearest Neighbors e Regressão Logística. | sim |
| [Wu 2009] | não | não | Híbrido de Floresta Aleatória e Redes Neurais | não |
| [Verbeke et al. 2012] | não | não | Métodos de Conjunto | não |
| [Zhang et al. 2012] | não | não | Híbrido envolvendo Floresta Aleatória, Regressão logística e Redes Neurais | não |
| [Chiang et al. 2003] | sim | não | Detecção de Padrões Sequenciais | não |
| [Coussement and Poel 2009] | não | sim | Regressão Logística, SVM e Floresta Aleatória | não |
| [Pimentel and Goldschmidt 2019] | sim | sim | Detecção de Padrões Sequenciais e Redes Neurais | não |

Além das pesquisas voltadas à predição de evasão no segmento educacional, existem diversos trabalhos em detecção antecipada de *churn* em outros segmentos. Como mostra a Tabela 1, em geral e de forma análoga aos trabalhos na área da Educação, eles também buscam aplicar algoritmos de aprendizado de máquina para construir modelos de classificação binária. Para tanto, utilizam informações sobre o perfil dos clientes, além de dados estatísticos consolidados sobre a relação entre cliente e empresa (e.g., quantidade de produtos adquiridos, gasto médio mensal, tempo de relacionamento, etc). A maioria desses trabalhos não considera nem os sentimentos registrados nas interações cliente-empresa e nem o aspecto temporal quanto à ocorrência dessas interações. São exceções os trabalhos apresentados em [Coussement and Poel 2009], [Chiang et al. 2003] e [Pimentel and Goldschmidt 2019]. [Coussement and Poel 2009] explora os sentimentos extraídos das interações entre o cliente e a empresa mas não considera o aspecto temporal de ocorrência das informações de sentimento, deixando, desta forma, de observar possíveis padrões que precedam a ocorrência de *churns*. [Chiang et al. 2003], por outro lado, considera a ordem cronológica e os tipos das interações cliente-empresa que precedem os eventos de *churn*, mas não os sentimentos advindos dos registros dessas

interações. Até onde foi possível observar, [Pimentel and Goldschmidt 2019] foi o único trabalho a considerar a combinação simultânea de informações sobre possíveis sentimentos presentes nas interações cliente-empresa com a ordem em que essas interações ocorrem ao longo do tempo. Cabe destacar, que, embora promissor, o referido trabalho foi aplicado apenas em dados da área de telecomunicações, fora do contexto educacional, foco de interesse do presente artigo.

Assim sendo, diante do levantamento do estado da arte realizado e resumido na Tabela 1, não foram identificadas pesquisas contendo evidências experimentais que validem ou refutem a hipótese levantada neste trabalho, justificando, portanto, a sua realização.

## 4. *SS-DetChurn*

Proposto em [Pimentel et al. 2019] e ilustrado na Figura 1, o método *SS-DetChurn* foi adaptado de forma a ser utilizado nos experimentos reportados neste artigo. Tal método requer a existência de $n \geq 1$ conjuntos de dados que contenham os registros históricos do conteúdo das interações realizadas entre a instituição de ensino e seus alunos por meio de canais de comunicação $C_1, C_2, ..., C_n$ tais como *E-mail*, *chat*, telefone, Portal *Online*, dentre outros. Também deve existir uma fonte de dados que indique a situação do aluno junto à instituição (i.e., *churn=sim*, no caso em que o aluno abandonou a instituição, ou *churn=não*, caso contrário). O método requer ainda que o conjunto de dados associado a cada $C_i$ possua, para cada interação, no mínimo, as seguintes informações: o momento em que a interação ocorreu, o aluno com o qual a interação ocorreu, o assunto que levou à interação e a transcrição da conversa registrada no momento da interação.

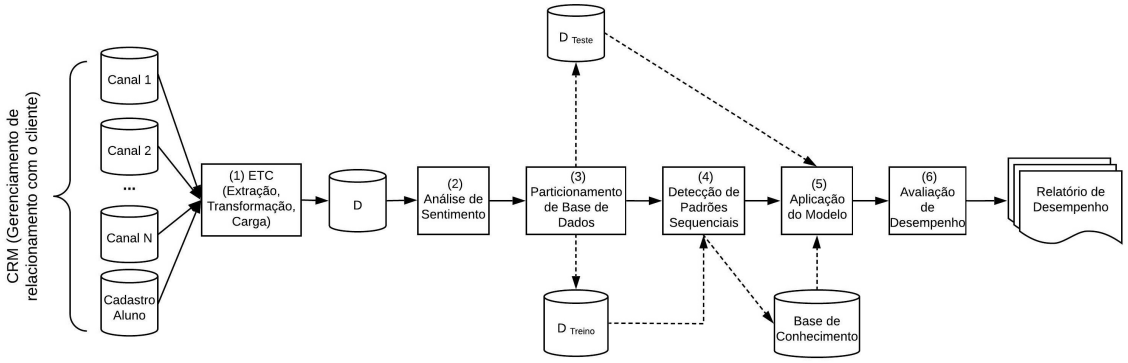


**Figura 1. Visão Macro-Funcional do *SS-DetChurn***

A etapa 1 (Extração, Transformação e Carga de Dados) é responsável por realizar todas as operações de seleção, adequação e integração dos dados que serão utilizados nas etapas seguintes do *SS-DetChurn*. Deve produzir como saída um conjunto de dados $D$ cuja estrutura está indicada no exemplo da Figura 2(a). Ela contém quatro atributos que indicam, para cada registro de interação $r$ ocorrida, a data $t$ de ocorrência de $r$, a identificação do aluno $o$ com o qual $r$ ocorreu, e o conjunto de itens $s$ associado a $r$. Convém notar que $s$ pode conter diversos itens tais como o canal $c$ que registrou $r$, o assunto $m$ que motivou a ocorrência de $r$, a transcrição da conversa $v$ registrada em $r$, dentre outros, incluindo a indicação de ocorrência de *churn*.

A informação sobre ocorrência ou não de *churn* deve ser obtida no conjunto de dados que contém o cadastro dos alunos. Assim, para cada aluno $o$, após a carga em $D$ de

todas as interações registradas nos canais de comunicação em que $o$ participou, inclui-se uma interação adicional $r$ vinculada a $o$ onde o conjunto de itens associado a $r$ contém apenas um item: *churn=sim* ou *churn=não*, retratando, assim, a situação de $o$ no cadastro de alunos. A data $t$ da interação $r$ recebe a data do desligamento informada no cadastro, caso o aluno tenha se desligado. Caso contrário, $t$ recebe a data de execução da etapa 1.

| Data | Aluno | Conjunto de Itens |
|---|---|---|
| 08/0617 | 1 | {canal=email; motivo=Problemas de acesso; Conversa="Prezado...} |
| 03/0717 | 1 | {canal=email; motivo=Reclamação sobre o curso; Conversa="Olá,...} |
| 05/07/17 | 1 | {canal=online; motivo=Problemas de acesso; Conversa="Prezado...} |
| 28/12/17 | 1 | {churn=sim} |
| 03/02/16 | 2 | {canal=online; motivo=Reclamação sobre o curso; Conversa="Caro... |
| 22/05/17 | 2 | {canal=telefone; motivo=Solicitação da 2a via senha; Conversa="Alô...} |
| 26/12/17 | 2 | {churn=sim} |
| 30/03/17 | 3 | {canal=chat; motivo= Solicitação da 2a via da senha; Conversa="Olá...} |
| 07/11/17 | 3 | {churn=não} |

(a)

| Conjunto de Itens após a etapa 2 |
|---|
| {canal=email; motivo=Problemas com acesso; Sentimento=Negativo} |
| {canal=email; motivo=Reclamação sobre o curso; Sentimento=Negativo} |
| {canal=online; motivo=Problemas com acesso; Sentimento=Negativo} |
| {churn=sim} |
| {canal=online; motivo=Reclamação sobre o curso; Sentimento=Negativo} |
| {canal=telefone; motivo=Solicitação da 2a via da senha; Sentimento=Positivo} |
| {churn=sim} |
| {canal=chat; motivo= Solicitação da 2a via da senha; Sentimento=Positivo} |
| {churn=não} |

(b)

**Figura 2. Exemplo de conjunto: (a) após a etapa 1; (b) após a etapa 2**

A etapa 2 (Análise de Sentimentos) consiste em aplicar, para cada registro $r$ em $D$, um algoritmo que seja capaz de classificar o texto da conversa $v$ associada a $r$ quanto à polaridade do sentimento subjacente a $v$: positiva ou negativa. Ao final do processo, o resultado da classificação da polaridade associada à $v$ é usado para atualizar o conjunto de itens $s$ de $r$ da seguinte forma: a polaridade identificada em $v$ é inserida em $s$, ao mesmo tempo em que $v$ é excluída de $s$. A Figura 2(b) mostra o conjunto de itens do conjunto de dados do exemplo da Figura 2(a) após a execução da etapa de análise de sentimentos. Cabe ressaltar que qualquer algoritmo de classificação de polaridade pode ser empregado nesta etapa e que a escolha desse algoritmo depende, fundamentalmente, da preferência do analista de dados responsável pela aplicação do *SS-DetChurn*.

A etapa 3 (Particionamento da Base de Dados) consiste em dividir $D$ aleatoriamente em dois subconjuntos, $D_{Treino}$ e $D_{Teste}$, tais que não existam alunos que possuam interações em $D_{Treino}$ e $D_{Teste}$ simultaneamente. Para tanto, o analista de dados deve especificar qual deve ser a proporção resultante de alunos nos dois subconjuntos. Um método de amostragem aleatória estratificada [Faceli et al. 2015] deve assegurar que a preservação da distribuição das classes *churn* e não *churn* existente em $D$.

Na etapa 4 (Detecção de Padrões Sequenciais), deve ser executado algum algoritmo de detecção de padrões sequenciais compatível com as definições apresentadas na Seção 2. Diante da diversidade de algoritmos desta natureza, cabe ao analista de dados optar por um deles. Em essência, o algoritmo escolhido precisa identificar padrões sequenciais $P = \ll p_1, p_2, ..., p_r \gg$ frequentes e válidos de tal forma que $\ll p_r \gg = \ll churn = sim \gg$ ou $\ll p_r \gg = \ll$ *churn=não* $\gg$. A escolha dos parâmetros de suporte e confiança mínimos também deve ser feita pelo analista de dados. O algoritmo escolhido deve ser aplicado sobre $D_{Treino}$. Após a identificação dos padrões sequenciais frequentes e válidos em $D_{Treino}$, esses são armazenados em uma base de conhecimento.

A etapa 5 (Aplicação do Modelo) tem como objetivo aplicar nos registros de dados de $D_{Teste}$ os padrões sequenciais minerados anteriormente. Assim, para cada aluno $o$ em $D_{Teste}$, verifica-se se existe alguma sequência $S_o = \ll s_1, s_2, ..., s_r \gg$ em $D_{Teste}$ e algum padrão sequencial $P = \ll p_1, p_2, ..., p_r \gg$ na base de conhecimento tal que $\ll p_1, p_2, ..., p_{r-1} \gg = \ll s_1, s_2, ..., s_{r-1} \gg$. Em caso positivo, compara-se se $\ll p_r \gg = \ll s_r \gg$. Caso a comparação seja verdadeira, é computado um acerto do modelo. Caso

a comparação seja falsa, computa-se um erro. Caso não exista padrão sequencial $P$ na base de conhecimento tal que $\ll p_1, p_2, ..., p_{r-1} \gg = \ll s_1, s_2, ..., s_{r-1} \gg$, assume-se *churn=não* a fim de realizar a comparação com $\ll s_r \gg$ e, então, computar erro/acerto.

Conforme o próprio nome sugere, a etapa 6 (Avaliação de Desempenho) tem como objetivo calcular alguma medida que expresse o grau de adequação do modelo gerado em detectar antecipadamente a ocorrência de *churn*. Tal medida deve ser função da quantidade de erros e acertos obtidos a partir da aplicação do modelo no conjunto $D_{Teste}$. *Precisão*, *Acurácia*, *Taxa de Falsos Positivos*, e *Área Sob a Curva*, entre outras, são exemplos de medidas popularmente utilizadas na avaliação de desempenho de problemas preditivos [Faceli et al. 2015]. Também neste ponto, a escolha cabe ao analista de dados.

## 5. Experimentos e resultados

Para avaliar a hipótese levantada neste trabalho, os experimentos foram realizados com os dados do *CRM* de uma universidade privada brasileira. Mais especificamente, foram considerados históricos de interação entre a universidade e seus alunos ocorridos ao longo dos doze meses do ano de 2017 por meio de quatro canais de comunicação: *E-mail*, *chat*, *voz* e Portal *Online*. A escolha deste cenário deveu-se basicamente à disponibilidade de acesso dos autores deste trabalho aos dados necessários aos experimentos[2]. A amostra utilizada continha 49.013 alunos com, no mínimo 1, no máximo 11, e, em média 6 interações. A distribuição dos dados foi de 87% para não *churn* e 13% para *churn*. A situação (*churn* /não *churn*) de cada aluno foi obtida do sistema de controle acadêmico da instituição.

Todos os experimentos realizados tomaram como base o método *SS-DetChurn* cujas etapas foram executadas com o apoio da ferramenta *SPSS Modeler*®.

Inicialmente, na etapa de *ETC*, foram realizadas operações de seleção e limpeza dos dados disponíveis nos diferentes canais de comunicação. Em seguida, os dados foram formatados e reunidos em uma tabela única. Cabe enfatizar que a existência de um cadastro único de alunos integrado aos bancos de dados dos diferentes canais evitou a necessidade de operações de duplicação de alunos para integração dos dados. Outro fator facilitador foi a existência de um conjunto único de motivos usado pelos canais para caracterizar de forma estruturada os assuntos tratados nas interações. Tal conjunto foi utilizado no enriquecimento dos conjuntos de itens das interações no momento de carga dos dados. Ao todo foram identificados vinte e quatro motivos de natureza administrativa, dezesseis de natureza acadêmica e três de ordem financeira, tais como recuperação de senha, correção notas e 2a via de boleto, respectivamente.

No momento seguinte à etapa de *ETC*, foi executada a etapa de análise de sentimentos. Nela, foi aplicado o algoritmo *NNLM feedforward* [Mikolov et al. 2013] sobre o texto da conversa de cada interação $r$ a fim de obter a polaridade do sentimento associado e enriquecer com tal informação o conjunto de itens de $r$. Cabe destacar que, para executar a classificação, o *NNLM feedforward* utiliza o *word2vec* na representação dos textos e uma rede neural recorrente treinada a partir de um corpus específico. Detalhes do treinamento da rede e dados estão em [Pimentel et al. 2019].

Na etapa de detecção de padrões sequenciais, foi utilizado o algoritmo BIDE

[2]No entanto, por questões de sigilo junto à IES, esses dados não puderam ser disponibilizados publicamente para *download*.

[Wang 2004] com os suporte e confiança mínimos de 30% e 80%, respectivamente, com valores experimentais. Tal algoritmo toma como base os conceitos apresentados na Seção 2 a fim de identificar padrões sequenciais frequentes e válidos no conjunto de dados.

A fim de permitir a comparação do efeito da utilização dos sentimentos na geração dos modelos preditivos, foram considerados dois cenários. No primeiro, o método adotado foi integralmente aplicado. Considerou-se, portanto, o uso dos sentimentos na detecção de padrões sequenciais. No segundo cenário, as informações de sentimento identificadas na etapa 3 não foram consideradas na construção dos modelos preditivos. Em ambos os cenários, a técnica de validação cruzada com *k-conjuntos* foi utilizada na avaliação dos modelos preditivos gerados. Duas métricas de avaliação foram calculadas: acurácia e taxa de falsos positivos. A Tabela 2 apresenta os desempenhos dos modelos obtidos na validação cruzada com $k = 5$. Observa-se que, de uma maneira geral, o modelo obtido sem considerar o atributo de sentimento, obteve um resultado significativo de 70,1% de acurácia. Entretanto, quando inserido o atributo de sentimento, o resultado aumentou em 14,5 p.p., obtendo uma acurácia de 84,6% que corrobora a hipótese inicial levantada neste trabalho. Estes números podem ser considerados como resultados satisfatórios, uma vez que a taxa de falsos positivos ficou em torno de 5,5%, valor inferior ao valor de 7,2% obtido diante do cenário de análise sem o atributo de sentimento.

**Tabela 2. Padrões sequenciais com maior frequência identificados na Etapa 3**

| Métrica de desempenho | Sem atributo de sentimento | Com atributo de sentimento |
|---|---|---|
| Taxa de Falsos Positivos | 7,2% | 5,5% |
| Acurácia Total | 70,1% | 84,6% |

Outro aspecto a ser comentado e que também aponta para a confirmação da hipótese investigada pode ser observado na Tabela 3 que contém os seis padrões sequenciais válidos com maior frequência identificada pelo algoritmo de mineração[3]. A tabela também indica a acurácia de cada um dos seis padrões em duas situações: uma em sua versão integral, com o item sentimento em cada conjunto de itens, e a outra em sua versão simplificada, onde o item sentimento foi removido de cada conjunto de itens. Pode-se perceber que a acurácia de cada padrão apresenta piora quando o atributo de polaridade de sentimento é removido, chegando a uma perda de 39,23% para o padrão 2. Tal fato reforça a importância do uso da informação sobre o sentimento associado a interações em ambientes educacionais na construção de modelos preditivos de evasão.

**Tabela 3. Padrões sequenciais com maior frequência identificados na Etapa 3**

| Descrição | Suporte | Confiança | Acurácia | | |
|---|---|---|---|---|---|
| | | | Sem atributo de sentimento | Com atributo de sentimento | Variação |
| (chat; nota; negativo); (email; solicitação de senha; negativo); (voz; ouvidoria; negativo) | 41,56% | 86,66% | 32,21% | 43,20% | 34,16% |
| (voz; ouvidoria; negativo); (voz, ouvidoria; negativo); (voz, ouvidoria; negativo) | 31,23% | 83,20% | 15,30% | 21,31% | 39,23% |
| (chat; nota; negativo); (email; nota; negativo); (voz; 2a via do boleto; negativo) | 34,21% | 80,10% | 9,35% | 11,35% | 21,39% |
| (email; 2a via boleto; negativo); (email; bilhete atrasado; negativo); (voz; ouvidoria; negativo) | 39,47% | 86,73% | 6,42% | 7,43% | 15,59% |
| (email; 2a via do boleto; negativo); (email; ouvidoria; negativo) | 47,57% | 85,91% | 2,29% | 2,89% | 26,20% |
| (voz; solicitação de senha; negativo); (email; solicitação de senha; negativo) | 30,28% | 89,33% | 1,58% | 1,78% | 12,66% |

É importante mencionar que, após uma análise qualitativa dos padrões indicados na Tabela 3, foi possível propor um conjunto de providências a serem incorporadas no

[3]Note que o item $churn = sim$ de cada padrão foi omitido por questão de simplificação, assim como o nome do atributo de cada item.

programa de combate à evasão escolar da instituição analisada. Os próximos parágrafos comentam as providências propostas mais relevantes.

Foi implementada uma melhoria do Portal do Aluno para incorporar uma funcionalidade que confere autonomia ao estudante para atualização de sua senha de acesso ao ambiente acadêmico online da instituição. Tal providência teve como objetivo mitigar os problemas relacionados à dificuldade na atualização de senhas de aluno, um dos mais recorrentes identificados pelos padrões minerados. Na mesma linha, foi incorporada ao Portal do Aluno uma funcionalidade que permite ao próprio estudante emitir segunda via de boletos de pagamento. Neste caso, o objetivo foi reduzir a insatisfação dos alunos decorrente do tempo de espera demandado pela instituição na emissão de segundas vias desses documentos. Também foi realizado o aprimoramento do *chat* institucional por meio da incorporação de *bots* inteligentes que pudessem auxiliar os alunos na resposta a questões frequentemente levantadas. Neste caso, os objetivos foram reduzir o tempo de espera do aluno e otimizar o tempo da equipe de atendimento, direcionando para ela apenas os casos envolvendo questões não triviais.

Duas outras providências foram a atualização da infraestrutura tecnológica do setor de ouvidoria institucional e o treinamento periódico da equipe de atendimento desse setor de forma a assegurar que o tempo de resposta para a solução dos problemas após os atendimentos ocorra dentro dos prazos previamente acordados com os alunos.

Outra providência foi o desenvolvimento de um sistema de monitoramento dos canais de comunicação do *CRM* baseado nos padrões minerados a fim de detectar novos alunos com potencial para evasão escolar. Em paralelo, foi criado um programa de treinamento periódico de uma equipe voltada à intervenção proativa junto aos casos de propensão à evasão detectados pelo sistema de monitoramento no dia a dia da instituição.

## 6. Considerações finais

O método proposto em *SS-DetChurn* foi adaptado para ser aplicado nos dados históricos de conversas entre alunos e representantes de uma instituição de ensino superior. Os experimentos realizados produziram resultados quantitativos que apontam para a validade da hipótese investigada e resultados qualitativos que permitiram a formulação de ações a serem incorporadas no programa de combate à evasão escolar da referida instituição. Entre os trabalhos futuros estão a avaliação dos efeitos das ações de combate à evasão elaboradas a partir dos resultados dos experimentos reportados e a comparação do *SS-DetChurn* com os métodos de pesquisas que modelam o problema de evasão usando perfil de aluno, além disso, uma análise estatística dos resultados evidenciando a hipótese que o uso de atributo de sentimento melhora o desempenho.

## Referências

Agrawal, R. et al. (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216.

Antonucci, D. (2018). 4 motivos para investir em um CRM especializado no ensino superior. Revista do Ensino Superior.

Bezerra, C. et al. (2016). Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes. CBIE.

Bocchini, B. (2018). Pesquisa sobre Evasão no Ensino Superior. Agência Brasil.

Chiang, D. A. et al. (2003). Goal-oriented sequential pattern for network banking churn analysis. volume 25, pages 293–302. ESWA.

Colpani, R. (2018). Mineração de Dados Educacionais: um Estudo da Evasão no Ensino Médio com Base nos Indicadores do Censo Escolar. volume 21. Informática na Educação: teoria prática.

Cordeiro, R. (2017). Identificação do Comportamento dos Estudantes Evadidos de Cursos Técnicos Utilizando Técnicas de Mineração de Dados. IFECTF - Dissert. Mestrado.

Coussement, K. and Poel, D. V. d. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. volume 36, pages 6127–6134. ESWA.

Faceli, K. et al. (2015). Inteligência Artificial. Uma Abordagem de Aprendizado de Máquina. LTC.

García, D. L. et al. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. volume 51, pages 719–774. KAIS.

Gonçalvez, T. C. et al. (2018). Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do IFMA. volume 10, pages 11–20. RBCA.

Hadden, J. et al. (2007). Computer assisted customer churn management: State-of-the-art and future trends. volume 34, pages 2902–2917. Computers and Operations Research.

Junior, J. G. O. et al. (2017). Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação. volume 13. Revista de Informática Aplicada.

Manhães, L. et al. (2012). Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados. SBSI.

Medeiros, L. and Padilha, T. (2018). Mineração de Dados para Detectar Evasão Escolar utilizando Algoritmos de Classificação: Um Estudo de Caso. CIET.

Mikolov, T. et al. (2013). Linguistic regularities in continuous space word representations. pages 746—-751. HLT-NAACL.

Pimentel, T. P. et al. (2019). Predição de churn baseada em detecção de padrões sequenciais e análise de sentimentos sobre as interações de clientes no CRM. IME - Dissert. Mestrado.

Pimentel, T. P. and Goldschmidt, R. R. (2019). Sequential Sentiment Pattern Mining to Predict Churn in CRM Systems: A Case Study with Telecom Data. volume 15, pages 1–8. SBSI'19 Proceedings of the XV Brazilian Symposium on Information Systems.

Ramos, J. et al. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. VII CBIE.

Verbeke, W. et al. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. volume 218, pages 211–229. EJOR.

Wang, J. (2004). BIDE: Efficient mining of frequent closed sequences. pages 79—90. IEEE Press.

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. volume 36, pages 9105–9112. ESWA.

Zhang, X. et al. (2012). Predicting customer churn through interpersonal influence. volume 28, pages 94–104. Knowledge-Based Systems.