UM ESTUDO ANALÍTICO SOBRE DADOS DE UMA CENTRAL DE ATENDIMENTO

André Philipe Andriotti de Moraes e Leandro Nunes de Castro
Apoio: PIBIETI Mackenzie

RESUMO

Em sua grande parte, as centrais de atendimento pouco ou não exploram os dados gerados durante as conversas, apesar de serem uma excelente base para melhorar o desempenho da empresa. Em decorrência disso, faz-se necessária a criação de uma plataforma computacional inteligente que objetiva a análise do humor do cliente e a identificação do assunto do contato, por meio de recursos analíticos como mineração de textos e processamento de linguagem natural. Para alcançar tais objetivos, foram construídas diversas aplicações na plataforma analítica SOMMA, da empresa AXONDATA Tecnologia Analítica LTDA, permitindo a classificação de sentimento por meio de processamento de texto, remoção de *stopwords*, geração de tokens, estruturação do documento em um *Bag of Words* e treinamento manual de um classificador. Não somente isso, mas também foi implementada uma aplicação focada na extração de palavras-chave, utilizando um componente responsável pela contagem de palavras em uma base de dados, o que possibilitou a geração de nuvens de palavras compostas pelas palavras mais frequentes no documento, trazendo um modelo mais visual dos resultados da aplicação. Como resultado, foi obtido um conjunto de informações sobre o atendimento, solucionando os objetivos propostos. Dessa forma, gerou-se um melhor aproveitamento dos dados proporcionados pela central de atendimento e maior satisfação dos clientes e atendentes.

Palavras-chave: Análise de dados. Mineração de textos. Central de atendimento.

ABSTRACT

For the most part, call centers use little or no data generated during conversations, despite being an excellent basis for improving company performance. As a result, it is necessary to create an intelligent computing platform that aims to analyze the customer's mood and identify the subject of the contact, through analytical resources such as text mining and natural language processing. To achieve these goals, several applications were built on the SOMMA analytical platform, from the company AXONDATA Tecnologia Analítica LTDA, allowing the classification of sentiment through text processing, stopword removal, token generation, document structuring in a Bag of Words and training of a classifier. Not only that, but an application focused on keyword extraction was also implemented, using a component responsible for counting words in a database, which enabled the generation of word clouds composed of the

most frequent words in the document, bringing a more visual model of the application's results. As a result, a set of information about the service is obtained, solving the proposed objectives. In this way, it generates a better use of the data provided by the call center and greater customer and attendant satisfaction.

Keywords: Data analysis. Text mining. Call center.

1. INTRODUÇÃO

As empresas nacionais e internacionais estão em constante evolução, em uma disputa acirrada para sobreviver e se destacar no mercado. Um pilar central desse processo é a qualidade do atendimento oferecido aos clientes e parceiros. Em decorrência disso, a criação das *centrais de atendimento* impulsionou diretamente o mercado de trabalho, uma vez que a chave para um negócio bem-sucedido é manter o consumidor satisfeito (GENEZE, 2017). As centrais de atendimento operam por meios digitais ou telefônicos, mantendo comunicação direta com os clientes, o que gera uma grande quantidade de dados textuais que podem ser analisados visando controle de qualidade e otimização operacional.

Nesse contexto, o objetivo do presente projeto é estudar, implementar e testar um conjunto de métodos analíticos para a extração de conhecimentos a partir de textos com aplicação a dados de centrais de atendimento. Como caso de estudo, serão utilizados dados de uma empresa parceira que possui uma central de atendimento a usuários com deficiência auditiva. Especificamente serão estudadas duas tarefas analíticas:

- **Identificação do objeto (assunto) do contato:** com mecanismos de identificação de contexto, reconhecimento de entidades e extração de palavras-chave, será possível identificar o assunto do contato. Essa funcionalidade é de suma importância para ranquear os motivos mais e menos recorrentes, possibilitando assim a criação de novas estratégias de mercado e até mesmo a correção de falhas do produto ou serviço.
- **Análise do humor do cliente:** por meio de técnicas de análise e classificação de sentimento, possibilitando definir o “humor” do cliente atendido. Em consequência disso, os operadores serão orientados intuitivamente a realizar ações e reparações para melhorar os atendimentos classificados como negativos e impulsionar os que se enquadram como positivos.

Tendo em vista os objetivos do projeto e os resultados da análise, será possível detectar, de forma rápida e automática questões como anseios, dificuldades e dúvidas dos clientes, proporcionado à empresa diversos benefícios, como maior eficiência no atendimento e satisfação dos clientes e atendentes.

2. REFERENCIAL TEÓRICO

2.1 Mineração de Textos

A Mineração de Textos (ou *Text Mining*) pode ser associada ao conceito de mineração, por meio da qual é possível extrair minerais preciosos. De maneira análoga, o processo

semi-automatizado para extração de conhecimento a partir de dados textuais não-estruturados e utilizando diferentes algoritmos é conhecido como *mineração de textos* (DE CASTRO; FERRARI, 2016a).

Apesar de os termos *Mineração de Textos* e *Mineração de Dados* objetivarem a identificação de padrões e informações relevantes, eles se diferem em relação à natureza do dado a ser analisado: dados não estruturados e estruturados, respectivamente. Estima-se que mais de 80% dos dados corporativos são não-estruturados (TERA, 2020), ou seja, dados encontrados em arquivos de texto, XML, documentos de Word, PDFs, entre outros modelos sem uma estrutura padrão, como imagens e gravações de áudio. Enquanto os dados não estruturados não possuem uma estrutura pré-definida, os dados estruturados são representados em tabelas de bancos de dados, onde não há a necessidade de realizar uma estruturação prévia da base para então aplicar os algoritmos de mineração e extrair insights de modo eficaz (HUPDATA, 2020).

Tendo em vista a grande quantidade de dados não estruturados em ambientes corporativos, principalmente dados textuais, a aplicação de algoritmos de mineração de textos apresenta-se como solução adequada para atender a alta demanda por análises desse tipo de dado, provando-se eficaz em diversas áreas. Como exemplo, podemos citar a Medicina, em que os dados gerados em ambientes hospitalares (fichas de pacientes, prontuários, exames, registros, etc.) podem auxiliar os profissionais da saúde a detectar doenças e providenciar o tratamento adequado. Não somente na área da saúde, mas também nos negócios, a mineração de textos pode ser usada na extração de informações de contratos, análise de sentimento em pesquisas de opinião, suporte e atendimento ao usuário. Essas são apenas algumas das mais variadas aplicações onde a mineração de textos exerce um papel essencial para o profissional, proporcionando maior eficiência e praticidade em seu trabalho (JUNIOR, 2008).

Devido à falta de padronização em documentos não estruturados, diferentes tipos de problemas podem ocorrer durante o processo de mineração, tais como: *incompletude*, na qual ocorre a falta de valores, atributos ou objetos nos dados; *inconsistência*, quando há valores muito discrepantes dos demais ou que não estejam no domínio do atributo; e *ruídos*, ou seja, variações indesejadas do valor de um dado, denominado *dado ruidoso* (DE CASTRO; FERRARI, 2016b). Portanto, faz-se necessária a preparação ou pré-processamento da base de dados a ser minerada de modo a evitar problemas na análise, sendo assim segmentada em três etapas principais:

- **Análise léxica ou *tokenização* do texto:** nesta etapa ocorre a conversão do documento (conjunto de cadeia de caracteres) em termos (*tokens*) com base em delimitações, como espaços, pontuações e fim de linha.
- **Remoção de *stopwords*:** etapa na qual palavras irrelevantes do texto (*stopwords*) são filtradas, tornando o documento mais claro e útil para a mineração. Esse processo é de extrema importância, devido ao caráter pouco discriminatório de palavras com alta frequência nos documentos.
- **Lematização (*stemming*):** por fim, é realizada a conversão dos termos de mesma raiz em um único formato padrão. Dessa maneira, são retiradas derivações, variações de gênero, sinônimos e plurais das palavras.

2.2 Análise de Sentimento

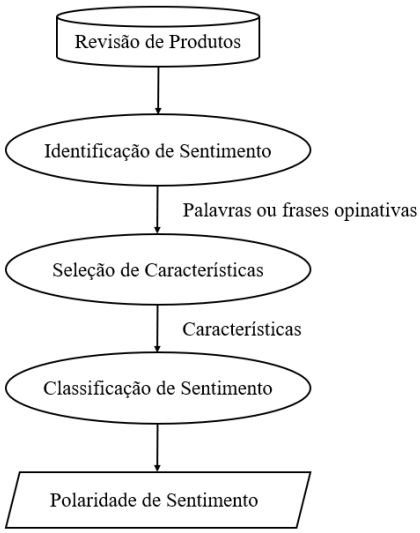
Análise de Sentimento (do inglês *Sentiment Analysis* (SA)) ou *Mineração de Opinião* (do inglês *Opinion Mining* (OM)) é o estudo computacional de opiniões, atitudes e emoções das pessoas, direcionadas a uma entidade, como produtos, organizações, indivíduos, eventos e seus atributos. Apesar de serem sinônimos, esses dois termos apresentam uma diferença: a OM extrai e analisa a opinião das pessoas sobre uma entidade, já a SA geralmente identifica o sentimento expresso em um texto e então o analisa (MEDHAT et al., 2014).

Por conta do explosivo crescimento no uso das mídias sociais (reviews, fóruns, blogs e redes sociais), o conteúdo gerado por essas ferramentas tem sido utilizado para tomar decisões, pois ele traz expressividade de opinião em seus dados. Exemplos de aplicações frequentes de SA incluem a revisão de produtos, mercado de ações, novos artigos e debates políticos (LIU; HIRST, 2012) (PANG; LEE, 2008).

De acordo com a pesquisa de Walaa Medhat, Ahmed Hassan e Hoda Korashy (2014) sobre algoritmos e aplicações de análise de sentimento, há três níveis de classificação em SA:

- **Nível do documento (*document-level*)**: classifica a opinião de todo o documento, expressando uma opinião/sentimento positivo ou negativo.
- **Nível da sentença (*sentence-level*)**: classifica o sentimento de cada sentença, identificando os trechos subjetivos e classificando-os em positivo ou negativo.
- **Nível de aspecto (*aspect-level*)**: classifica o sentimento de forma mais detalhada com base nos aspectos da entidade, identificando as entidades e seus aspectos.

Figura 1. Fluxograma de processos da análise de sentimento.



Fonte. Adaptado de Ain Shams Engineering Journal (MEDHAT et al., 2014).

A *seleção de características* (do inglês *Feature Selection* (FS)) e a *classificação de sentimento* (do inglês *Sentiment Classification* (SC)) são dois campos muito discutidos entre pesquisadores, pois, nos últimos anos, diversas aplicações e aprimoramento para algoritmos de SA foram e estão sendo desenvolvidos com base em FS e SC. Dentre as áreas pesquisadas, encontram-se: *detecção de emoção* (do inglês *Emotion Detection* (ED)), cujo foco é extrair e analisar sentimentos explícitos ou implícitos nas sentenças, *construção de recursos* (do inglês *Building Resources* (BR)), a qual busca criar uma léxica, relatando as expressões de opinião com base em sua polaridade e em dicionários, e *transferência de aprendizagem* (do inglês *Transfer Learning* (TL)) ou *classificação entre domínios* (do inglês *Cross-Domain classification*), que analisa dados de um domínio e aplica os resultados em outro domínio específico (MEDHAT et al., 2014).

2.3 Seleção de Características

Segundo Medhat (2014), o primeiro passo para classificar sentimentos é a extração e seleção de características (*features*) do texto, como, por exemplo: presença ou não do termo (0 ou 1); frequência; classes gramaticais (principalmente adjetivos); sintaxe; frases e palavras usadas para expressar uma opinião; e negações.

Os métodos de seleção de características são divididos em dois tipos principais (MEDHAT et al., 2014). Baseados em léxico (*lexicon-based*), que utilizam um pequeno conjunto de palavras primitivas, seguido de um detector de sinônimos ou recursos léxicos

online, portanto requerendo supervisão humana. O outro tipo é composto pelos métodos estatísticos (*statistical methods*), totalmente automáticos, logo são mais utilizados. Dentre as etapas de FS, destacam-se a remoção de stopwords e de derivações (*stemming*), geralmente aplicadas em documentos *Bag of Words* (BoWs), devido sua simplicidade.

Tendo em vista esses conceitos de FS, três métodos estatísticos se destacam devido sua eficiência e simplicidade:

- ***Point-wise Mutual Information (PMI):*** O método estatístico PMI calcula o nível da relação entre a co-ocorrência esperada de uma classe *i* e uma palavra *w* (P_i • F(*w*)) e a sua co-ocorrência verdadeira (F(*w*) • p_i(*w*)). Portanto, se houver relação, $M_i > 0$, caso contrário, $M_i < 0$, com base na seguinte fórmula:

$$M_i(w) = \log\left(\frac{F(w) \cdot p_i(w)}{F(w) \cdot P_i}\right) = \log\left(\frac{p_i(w)}{P_i}\right)$$

  Esse método foi aprimorado por Yu e Wu (2013), acrescentando a distribuição contextual das palavras como um fator para identificar paravras mais significativas para as emoções, por meio de um modelo de entropia contextual que busca a similaridade entre palavras.

- ***Chi-square (χ^2):*** Assim como a PMI, o χ^2 calcula a correlação entre palavra e classe, porém é considerado um método melhor, pois apresenta valores normalizados, ou seja, são mais comparáveis entre palavras de mesma classe. A expressão que define esse método é:

$$\chi_i^2 = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot \big(1 - F(w)\big) \cdot P_i \cdot (1 - P_i)}$$

  onde *n* é o número de documentos na coleção, P_i é a fração global de documentos contendo a classe *i*, *F*(*w*) é a fração global de documentos contendo a palavra *w*, e p_i(*w*) é a probabilidade condicional da classe *i* para documentos contendo a palavra *w*.

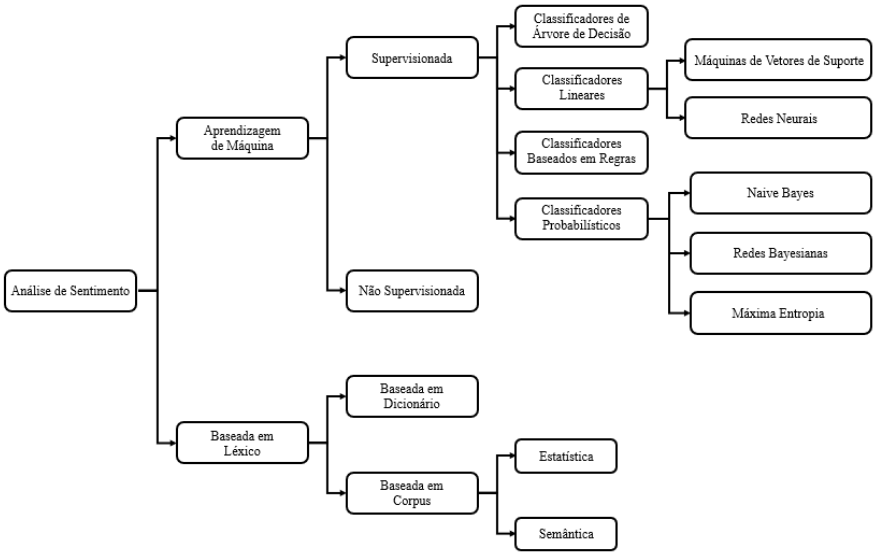
2.4 Classificação de Sentimento

Com base na pesquisa de Medhat W. et al. (2014), o campo de classificação de sentimento tem por objetivo classificar a opinião de um documento, expressando um

sentimento/opinião positivo ou negativo (nível do documento), assumindo que pertence a uma única entidade. Logo, se a opinião pertencer a mais de uma entidade, então o sentimento na entidade pode assumir um significado diferente.

As técnicas de SC podem ser divididas em três ramos, com suas respectivas subdivisões (MEDHAT et al., 2014): *aprendizagem de máquina* (subdividido em supervisionada e não supervisionada), *baseada em léxico* (subdividida em baseada em dicionário e baseada em corpus) e *híbrida* (combinação das anteriores).

Figura 2. Técnicas de classificação de sentimento.



Fonte: Ain Shams Engineering Journal (MEDHAT et al., 2014).

A classificação de sentimento é essencialmente um problema de classificação de texto, que geralmente classifica documentos em tópicos diferentes, como política, ciência, esporte, etc., onde as palavras relacionadas a esses tópicos são as principais características. Porém, em classificação de sentimentos, palavras que indicam opiniões e sentimentos positivos ou negativos são mais importantes, tais como “ótimo”, “péssimo”, “incrível” e “horrível” (LIU; HIRST, 2012). Portanto, desde que seja um problema de classificação de texto e que os documentos de treino estejam rotulados, qualquer método de aprendizagem supervisionada pode ser aplicado.

Os modelos que consideram probabilidades (ou distribuição de probabilidades), na estimação da classificação recebem o nome de *classificadores probabilísticos*, conhecido como *classificador generativo*. Dentre os classificadores probabilísticos mais conhecidos destaca-se o classificador *Naïve Bayes* (NB). O NB está entre os classificadores mais

simples e usados. Com ele é possível prever a probabilidade de um objeto pertencer a determinada classe (rótulo) a partir do Teorema de Bayes:

$$P(label|features) = \frac{P(label) \cdot P(features|label)}{P(features)}$$

onde *P*(*label*) é a probabilidade de um rótulo ou de uma característica aleatória definir um rótulo, *P*(*label*|*features*) é a probabilidade de um conjunto de características ser classificado como um rótulo, e *P*(*features*) é a probabilidade de haver um conjunto de características (MEDHAT et al., 2014).

3. METODOLOGIA

Para o desenvolvimento do projeto, foi selecionada uma base de dados que permitiu treinar um classificador capaz de distinguir o sentimento (humor) do cliente e do atendente em positivo e negativo de acordo com as palavras do diálogo entre eles durante o atendimento. Essa base de dados foi fornecida por uma empresa parceira e contém os registros dos atendimentos de uma central de atendimento direcionada a deficientes auditivos, com um total de 1.730 registros capturados durante o mês de julho de 2020. Para a análise desses dados foram construídas aplicações na plataforma analítica SOMMA.

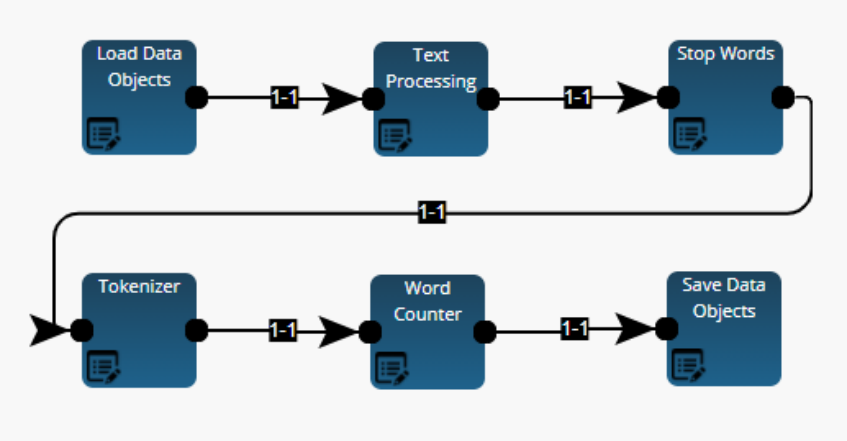
4.1 Identificação do Objeto do Contato

A primeira aplicação construída na SOMMA teve como objetivo a identificação automática do objeto (assunto) do contato e foi baseada em um método simples de análise de frequência de palavras e representação via nuvens de palavras (*tag clouds*).

Este método utiliza os processos de estruturação de documentos apresentados na Seção 2.1 para identificar as palavras do documento e depois faz a contagem da frequência de ocorrência de cada palavra nos documentos. A partir dessa contagem as palavras são ranqueadas em ordem decrescente e uma nuvem de palavras é plotada indicando aquelas de maior frequência. Além disso, podem ser feitas combinações de mais de uma palavra, formando o que chamamos de *n*-gramas, sendo *n* o número de palavras combinadas em sequência. Os *n*-gramas nos permitem ter uma melhor compreensão do contexto, pois trazem palavras que coocorrem em sequência nos documentos.

A Figura 3 ilustra a aplicação SOMMA construída para a contagem de *n*-gramas e extração de palavras-chave dos documentos. O componente *text processing* faz o processamento dos textos removendo acentos e símbolos, e transformando todas as letras para minúsculo. Em seguida as *stopwords* são removidas, os textos são toquenizados e é determinada a frequência absoluta de ocorrência das palavras resultantes.

Figura 3. Aplicação na plataforma SOMMA para o pré-processamento da base textual e contagem de palavras.



Após a execução da aplicação, foram geradas três tabelas para cada pessoa no diálogo (atendente e cliente), com os unigramas, bigramas e trigramas mais frequentes no documento e suas respectivas frequências de ocorrência. Entretanto, a existência de palavras irrelevantes era extremamente prejudicial para a extração de palavras-chave, devido a sua incapacidade de trazer informação útil à análise. Por isso, tais palavras foram removidas manualmente dos conjuntos de *n*-gramas mais frequentes, para que os documentos pudessem ser devidamente modelados em seis nuvens de palavras (ou *tagclouds*), obtendo assim uma visão mais clara das tabelas referidas.

As Figuras 4 a 6 apresentam as nuvens de palavras para os unigramas, bigramas e trigramas dos atendentes, respectivamente, e as Figuras 7 a 9 apresentam as nuvens de palavras para os unigramas, bigramas e trigramas dos surdos, respectivamente. Diante disso, é possível identificar automaticamente os principais assuntos discutidos pelos atendentes e surdos por meio do ranqueamento dos unigramas, bigramas e trigramas, com base na frequência de ocorrência desses termos, como pode ser observado nas Tabelas 1 a 6.

Figura 4. Nuvem de palavras composta pelos unigramas mais frequentes nas falas do atendente.

Tabela 1. Ranqueamento dos cinco unigramas mais frequentes nas falas do atendente e suas respectivas ocorrências.

Posição	Unigramas	Ocorrências
1º	numero	923
2º	libras	585
3º	telefone	310
4º	cpf	266
5º	contato	246

Figura 5. Nuvem de palavras composta pelos bigramas mais frequentes nas falas do atendente.

Tabela 2. Ranqueamento dos cinco bigramas mais frequentes nas falas do atendente e suas respectivas ocorrências.

Posição	Bigramas	Ocorrências
1º	central libras	549
2º	vídeo chamada	65
3º	icom gente	58
4º	outro interprete	26
5º	telefone fixo	26

Figura 6. Nuvem de palavras composta pelos trigramas mais frequentes nas falas do atendente.

Tabela 3. Ranqueamento dos cinco trigramas mais frequentes nas falas do atendente e suas respectivas ocorrências.

Posição	Trigramas	Ocorrências
1º	central libras gente	443
2º	por falta comunicação	69
3º	icom gente fala	58
4º	chamando caixa postal	16
5º	vivo sou atendente	11

Figura 7. Nuvem de palavras composta pelos unigramas mais frequentes nas falas do surdo.

Tabela 4. Ranqueamento dos cinco unigramas mais frequentes nas falas do surdo e suas respectivas ocorrências.

Posição	Unigramas	Ocorrências
1º	numero	524
2º	internet	251
3º	vivo	237
4º	telefone	233
5º	libras	176

Figura 8. Nuvem de palavras composta pelos bigramas mais frequentes nas falas do surdo.

Tabela 5. Ranqueamento dos cinco bigramas mais frequentes nas falas do surdo e suas respectivas ocorrências.

Posição	Bigramas	Ocorrências
1º	central libras	55
2º	meu numero	51
3º	meu filho	47
4º	ligar vivo	38
5º	vivo fixo	33

Figura 9. Nuvem de palavras composta pelos trigramas mais frequentes nas falas do surdo.

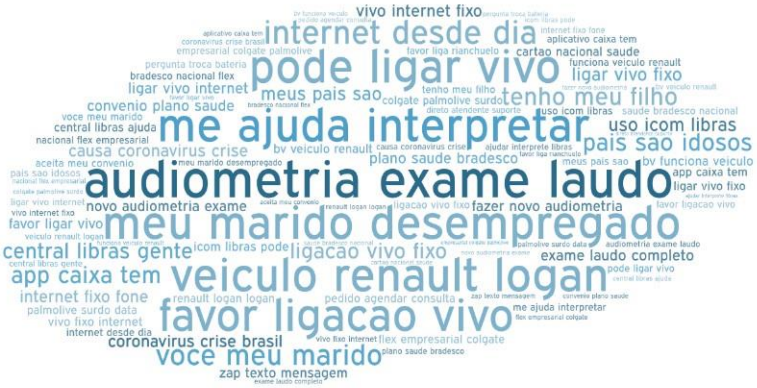


Tabela 6. Ranqueamento dos cinco trigramas mais frequentes nas falas do surdo e suas respectivas ocorrências.

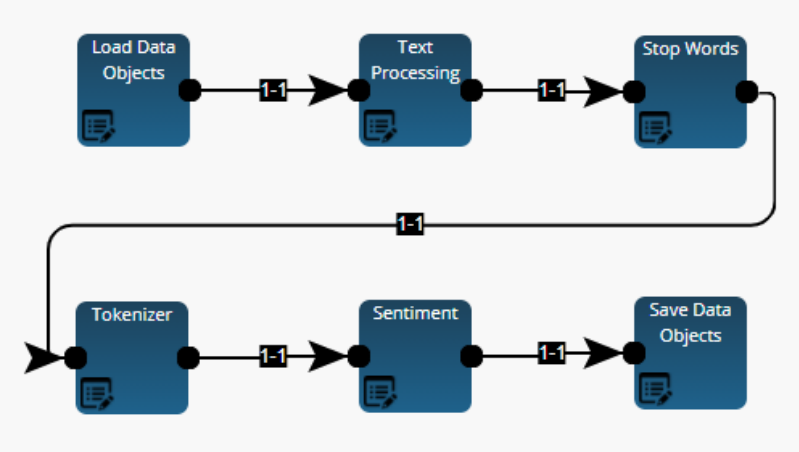
Posição	Trigramas	Ocorrências
1º	audiometria exame laudo	19
2º	meu marido desempregado	17
3º	me ajuda interpretar	16
4º	veiculo renaut logan	12
5º	pode ligar vivo	12

4.2 Classificação de Sentimento

Neste projeto a classificação de sentimento será realizada utilizando duas técnicas distintas, um dicionário (léxico) e um classificador Bayesiano, conforme discutido na Seção 2.4.

No primeiro caso (léxico), foi estruturada uma aplicação SOMMA para a preparação dos dados e a análise e classificação de sentimento utilizando o próprio classificador da plataforma, o componente *Sentiment*, retornando assim os atendimentos segmentados em positivos, negativos e neutros (null) (Figura 10).

Figura 10. Aplicação na plataforma SOMMA para o pré-processamento da base e a análise de sentimento utilizando o componente “Sentiment”, que é baseado em dicionário.



Dessa forma, obteve-se dois conjuntos de resultados (classificações dos atendentes e dos surdos), os quais são representadas nas Figuras 13 e 14. Ao analisar os dados, observou-se que 676 atendimentos do atendente e 813 atendimentos do surdo apresentavam dados nulos e, ao realizar a união entre os casos, obteve-se um total de 868 atendimentos considerados nulos dos 1730 (em sua maioria, devido à falta de comunicação com o cliente), ou seja, são informações que não são necessárias para a análise. Portanto, esses dados foram excluídos da base de análise para que o estudo pudesse focar apenas nos casos positivo ou negativo.

De modo semelhante, foi desenvolvida uma aplicação na SOMMA cujo objetivo é treinar um classificador (componente “*Classification*”) com a base de dados original tendo alguns de seus registros rotulados manualmente considerando as classes positivo e negativo (Figura 11). Neste caso, 100 registros foram selecionados aleatoriamente e rotulados manualmente para treinar o classificador. Do total de 100 registros do atendente, 80 foram

classificados como positivos e, para o caso dos surdos, 70 foram classificados como positivos, conforme mostra a Figura 12.

Figura 11. Aplicação na plataforma SOMMA para o pré-processamento da base de dados e a análise de sentimento com o classificador “Classification” treinado manualmente.

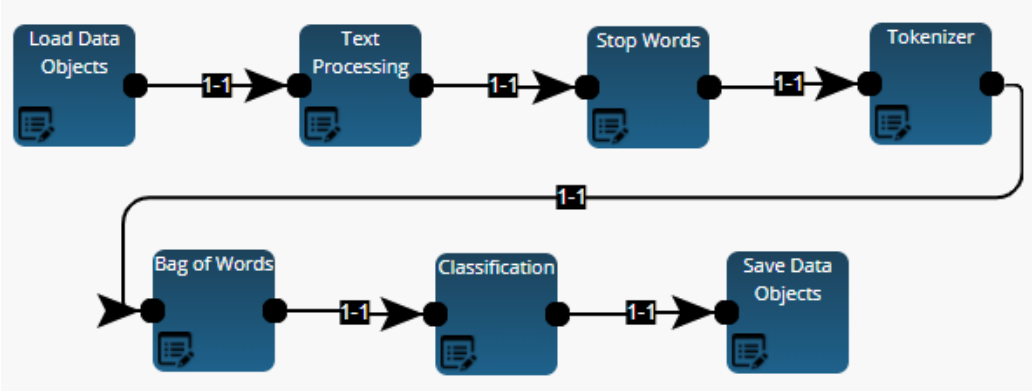
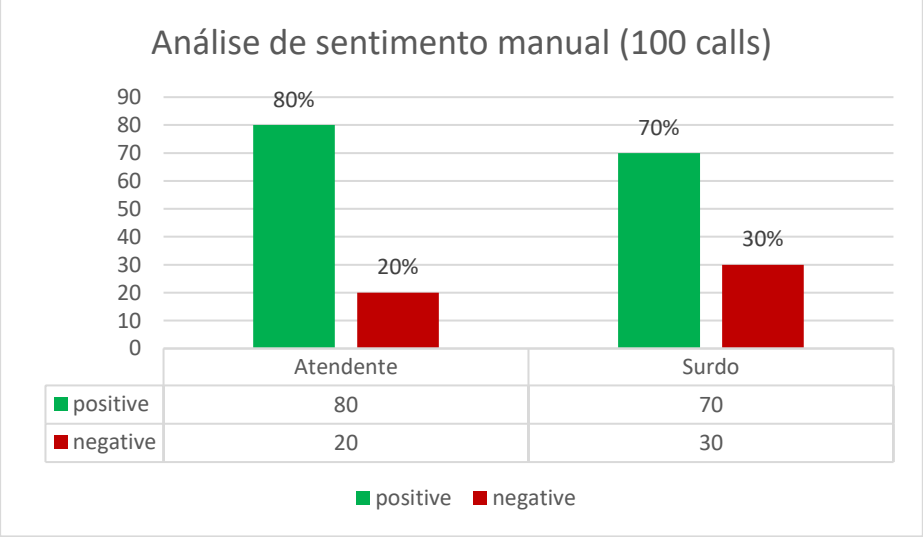


Figura 12. Base de dados classificada manualmente para o treinamento do classificador.



Após treinar o classificador, executou-se a aplicação com a base de dados original (1730 atendimentos) e com a base a qual foram retirados os diálogos classificados como nulos (862 atendimentos). Vale destacar que diferentes tipos de classificadores foram aplicados para a análise, porém os algoritmos mais eficazes para a classificação do sentimento do atendente e surdo foram respectivamente *Gradient Boost* e *Naive Bayes*, cujos detalhes e estatísticas se encontram na Tabela 7.

Tabela 7. Classificador mais eficaz para a classificação de sentimento do atendente e surdo e seus respectivos desempenhos.

Atendente				**Surdo**			
Classificador: Gradient Boost				Classificador: Naive Bayes			
Melhor f1 = 0.789474 (em dados de validação)				Melhor f1 = 0.777778 (em dados de validação)			
Métricas por classe:				Métricas por classe:			
	Positivo	Negativo	Peso		Positivo	Negativo	Peso
F1	0.981	0.926	0.97	F1	0.884	0.679	0.82
Precisão	0.987	0.904	0.97	Precisão	0.844	0.782	0.82
Revocação	0.975	0.95	0.97	Revocação	0.928	0.6	0.83
TPR	0.975	0.95	0.97	TPR	0.928	0.6	0.83
FPR	0.05	0.025	0.04	FPR	0.4	0.071	0.30
Acurácia	-	-	0.97	Acurácia	-	-	0.83
Matriz de Confusão:				Matriz de Confusão:			
	Positivo	Negativo			Positivo	Negativo	
Positivo	78.0	2.0		Positivo	65.0	5.0	
Negativo	1.0	19.0		Negativo	12.0	18.0	

4. RESULTADO E DISCUSSÃO

Ao analisar as nuvens de palavras e suas respectivas tabelas com o ranqueamento dos termos mais frequentes do diálogo entre o surdo e o atendente, infere-se que assuntos como problemas de internet, operadoras telefônicas (principalmente a Vivo), saúde, complicações familiares e de comunicação (com ênfase aos casos em que o cliente solicita um intérprete de Libras) apresentam uma alta frequência de ocorrência em comparação aos demais contextos, ou seja, são as adversidades mais presentes no cotidiano de um deficiente auditivo, que, por meio do auxílio das centrais de atendimento, podem ser solucionadas de forma prática e eficiente.

Em relação à classificação de sentimento feita através do componente "*Sentiment*" (baseado em dicionário), observarmos que, nos dados da Figura 13, a maior parte dos registros é classificada como positivo, principalmente para os atendentes. A Figura 14, que considera apenas os registros com sentimento não nulo, mostra essa predominância clara entre o sentimento positivo frente ao negativo.

Figura 13. Análise de sentimento com o classificador “Sentiment” da SOMMA (base de dados completa).

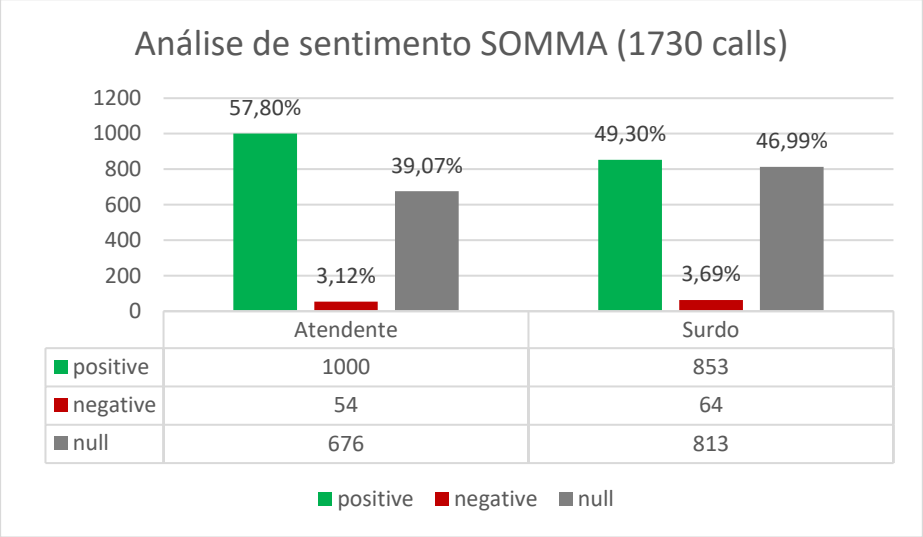
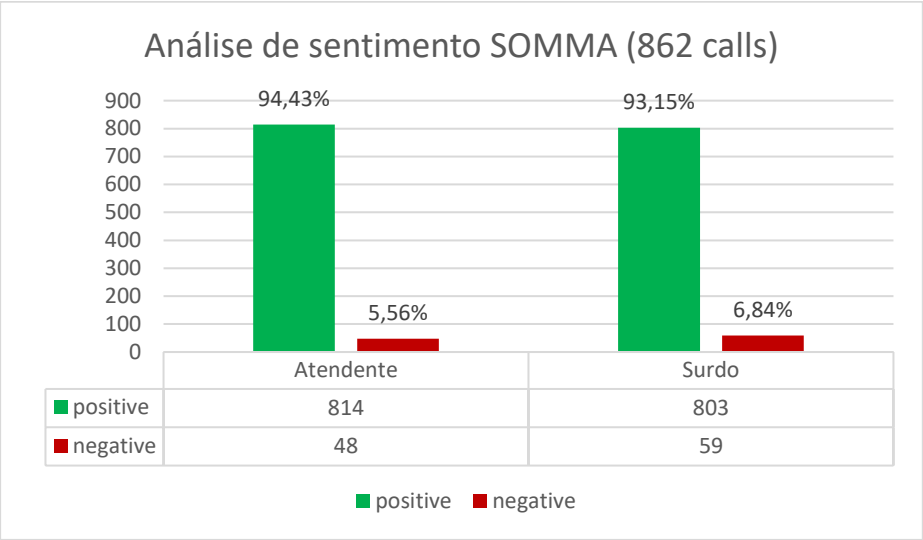


Figura 14. Análise de sentimento com o classificador “Sentiment” da SOMMA (base de dados após a remoção dos registros com sentimento “*null*”).



Após o treinamento do classificador rotulado manualmente (componente *“Classifier”* da SOMMA), pôde-se perceber a similaridade nos dados gerados aos dados resultantes da análise de sentimento com o componente *“Sentiment”*, como pode ser observado nas Figuras 15 e 16. Mesmo sem ou com a remoção dos casos com sentimento *null* (1730 atendimentos e 862 atendimentos, respectivamente), a predominância da classificação *positive* é significativa, com uma média de casos considerados como sentimento positivo igual a 87,43%, demasiadamente maior que os casos de sentimento negativo, cuja média é de 12,57%.

Figura 15. Análise de sentimento com o classificador “Classification” da SOMMA treinado manualmente (base de dados completa).

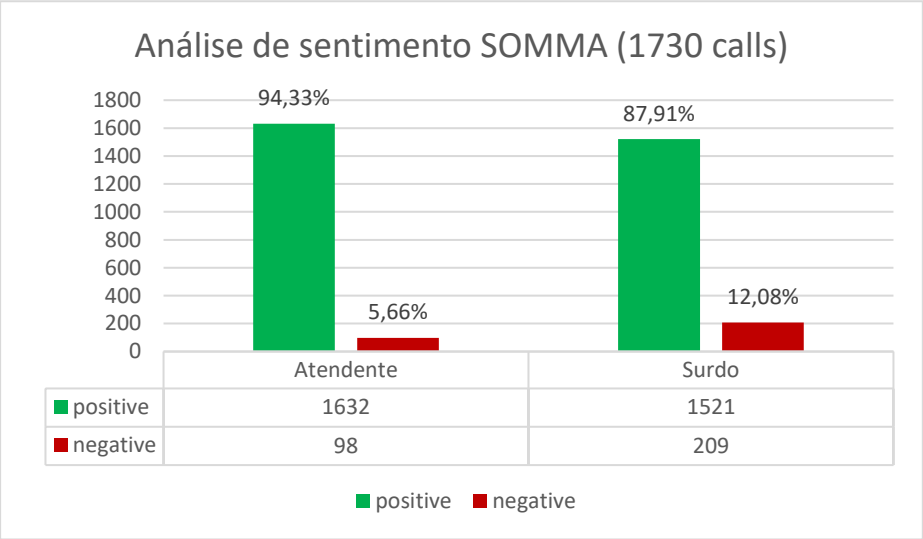
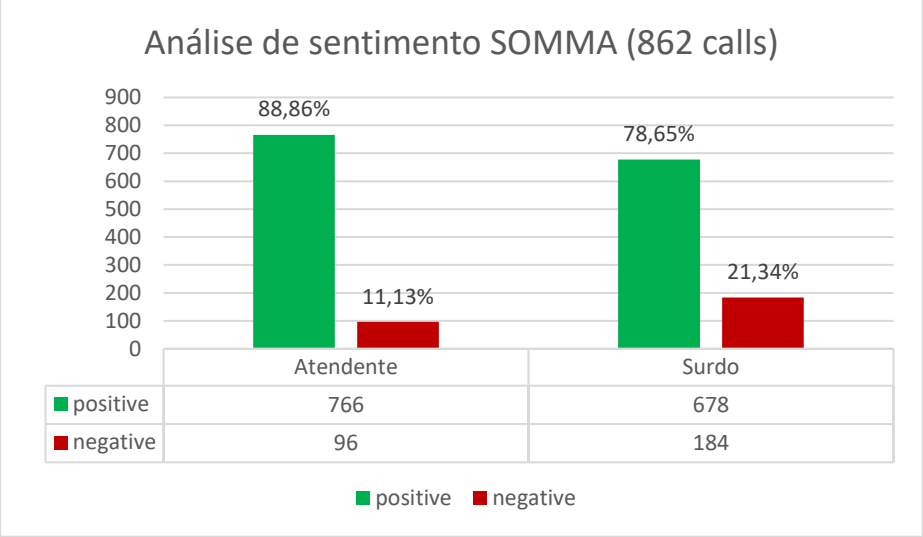


Figura 16. Análise de sentimento com o classificador “Classification” da SOMMA treinado manualmente (base de dados após a remoção dos registros com sentimento “null”).



5. CONSIDERAÇÕES FINAIS

Neste artigo realizamos um estudo analítico sobre dados de uma central de atendimento por meio da mineração de textos, processamento de linguagem natural e extração de palavras-chave com os seguintes objetivos principais: identificar o objeto (assunto) pelo qual o cliente recorreu ao atendente e analisar o humor do cliente.

Diante de tais objetivos, geramos aplicações na plataforma analítica SOMMA, onde realizamos o pré-processamento da base de dados (com processamento de texto, remoção de *stopwords*, tokenização e reestruturação do documento em um *Bag of Words*), para que houvesse um melhor proveito dos dados na análise. Após o pré-processamento, aplicamos

um contador de palavras para identificar os termos mais frequentes no diálogo, obtendo assim o assunto do atendimento, apresentado de forma mais clara em *TagClouds* e tabelas na Seção 4.1. Já em relação à análise de sentimento, utilizamos dois tipos de classificadores: o componente "*Sentiment*", um classificador de sentimento da própria SOMMA, e o componente "*Classification*", que foi treinado por uma amostra aleatória da base original, rotulada manualmente. Diante dessas análises, pôde-se classificar o sentimento do atendente e do surdo, que em grande parte foram classificados como positivos por ambos os classificadores, concluindo assim o segundo objetivo do projeto.

AGRADECIMENTOS

O aluno pesquisador agradece ao Santander pela bolsa de Iniciação Tecnológica, à Universidade Presbiteriana Mackenzie pela oportunidade e ao CNPq, Fapesp e MackPesquisa pelo apoio financeiro.

6. REFERÊNCIAS

B. Liu e G. Hirst, Sentiment Analysis and Opinion Mining: Synesthesis Lectures on Human Language Technologies, University of Illinois at Chicago: Morgan & Claypool Publishers, 2012.

B. Pang e L. Lee, Opinion mining and sentiment analysis, Estados Unidos: NOW, 2008.

HupData Administrator, "Mineração de Texto - O que é? Como aplicar?," HupData - Data Analysis Solutions, Maio 2020. [Online]. Available: https://hupdata.com/mineracao-de-texto-o-que-e-como-aplicar/. [Acesso em 3 Julho 2021].

J. R. Carrilho Junior, "Desenvolvimento de uma Metodologia para Mineração de Textos," Rio de Janeiro, 2008.

L.-C. Yu, J.-L. Wu, P.-C. Chang e H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," Knowledge-Based Systems, 2013. [Online]. Available: https://doi.org/10.1016/j.knosys.2013.01.001. [Acesso em 3 Julho 2021].

L. N. de Castro e D. G. Ferrari, "Capítulo 1 - Introdução à mineração de dados," em Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações, São Paulo, Saraiva, 2016, pp. 1-23.

L. N. de Castro e D. G. Ferrari, "Capítulo 2 - Pré-processamento de dados," em Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações, São Paulo, Saraiva, 2016, pp. 25-58.

P. Geneze, “O que é Service Desk?,” NeoAssist, Junho 2017. [Online]. Available: https://www.neoassist.com/blog/o-que-e-service-desk/. [Acesso em 3 Julho 2021].

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer e R. Harshman, “Indexing by latent semantic analysis,” JASIS, vol. 41, nº 6, pp. 391-407, 1990.

Tera, “O grande problema de lidar com dados não estruturados,” Medium, Agosto 2020. [Online]. Available: https://medium.com/somos-tera/o-grande-problema-de-lidar-com-dados-nao-estruturados-a307fb4f3200. [Acesso em 3 Julho 2021].

W. Medhat, A. Hassan e H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, Maio 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2090447914000550. [Acesso em 3 Julho 2021].

Contatos: andre.andriotti@hotmail.com e leandro.silva1@mackenzie.br