UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DIEGO DE SOUZA SILVA

OTTO: DESCOBRINDO ASSUNTOS E SENTIMENTOS EM COMENTÁRIOS

SÃO LEOPOLDO
2019

Diego de Souza Silva

OTTO: Descobrindo assuntos e sentimentos em comentários

Artigo apresentado como requisito parcial para obtenção do título de Bacharel em Ciência da Computação, pelo Curso de Ciência da Computação da Universidade do Vale do Rio dos Sinos - UNISINOS

Orientador: Prof. Dr. Mateus Raeder

São Leopoldo

2019

OTTO: DESCOBRINDO ASSUNTOS E SENTIMENTOS EM COMENTÁRIOS

Diego de Souza Silva*

Mateus Raeder**

Resumo: Este artigo apresenta um breve estudo sobre técnicas de mineração de texto e explora suas aplicações em comentários em português. Foi proposto um modelo capaz de analisar textos com a extração de assuntos e detecção do sentimento do autor. O modelo proposto foi aplicado em avaliações de restaurantes capturadas no TripAdvisor. LDA foi utilizado para modelagem de tópicos, enquanto SVM e Naive Bayes foram usados para a análise de sentimentos. Para fins comparativos, se propõe uma abordagem simples que utiliza dicionário de palavras e léxico. O método baseado em aprendizagem de máquina foi capaz de identificar assuntos e sentimentos e obteve melhores resultados do que a abordagem simples. Porém, a análise de sentimentos apresentou dificuldades em identificar textos negativos quando testado com usuários reais, problema que pode ser contornado com a aplicação de hiper-parâmetros e uma base de treinamento maior. O artigo conclui que o modelo proposto pode ser utilizado em sistemas de tomada de decisões.

Palavras-chave: mineração de texto, análise de sentimentos, classificação de texto, extração de tópico, modelagem de tópicos, lda.

1 INTRODUÇÃO

A opinião de clientes a respeito de produtos ou serviços adquiridos nunca foi tão importante quanto nos dias de hoje. Um bom relacionamento com o cliente pode levar tanto ao sucesso quanto ao fracasso de vendas. As redes sociais aproximaram ainda mais consumidores e empresas. Através delas, clientes podem realizar avaliações nas quais expõem problemas, descrevem serviços, relatam o atendimento, exaltam qualidades. Ou seja, descrevem toda a experiência de relacionamento com a empresa.

A literatura conta com diversos estudos relacionados às interações *online* entre clientes e empresas. Salinca (2015) alega que milhões de pessoas expressam seus pensamentos sobre vários produtos ou serviços nas redes sociais, blogs ou sites de avaliações populares. O *feedback* ativo das pessoas é valioso, não apenas para as empresas analisarem a satisfação de seus clientes e o monitoramento dos concorrentes, mas também é muito útil para os consumidores que desejam pesquisar

* Aluno do curso de Ciência da Computação na Unisinos. E-mail: diegoss.hhh@gmail.com.
** Doutor em Ciência da Computação. Coordenador do curso de Ciência da Computação na Unisinos. E-mail: mraeder@unisinos.br.

um produto ou serviço antes de efetuar uma compra. Sarakit et al. (2015) afirmam que as mídias sociais atuais possuem um importante papel no nosso dia a dia e nos negócios, especialmente promovendo produtos e atividades de marketing online. Sharma, Nigam e Jain (2014) afirmam que um grande número de comentários e sugestões de usuários está presente na *web* nos dias de hoje. As avaliações podem ser sobre produtos, críticas sobre filmes, entre outras - o que ajuda outros usuários a tomarem decisões.

As avaliações de clientes estão disponíveis a outros consumidores na internet. Dessa forma, novos clientes podem levá-las em consideração antes de comprar, ou não, em uma determinada empresa. Tal fato é um indicativo de que a interação *online* é um fator muito importante para a conversão de clientes. Sendo assim, empresas devem estar muito atentas às opiniões recebidas, lendo e respondendo adequadamente, de forma a criar um ambiente confiável para o consumidor.

Entretanto, o volume de avaliações recebidas por vezes é enorme, e dar atenção a todos os comentários pode acabar se tornando uma tarefa muito difícil. Prabowo e Purwarianti (2017), em sua pesquisa sobre lojas virtuais no Instagram afirmam que é difícil para o vendedor ler cada postagem antes de respondê-las. Seria ótimo se houvesse um sistema que pudesse classificar os comentários do Instagram com base na melhor resposta a ser dada. Segundo Zvarevashe e Olugbara (2018), nos últimos anos o mundo experimentou um grande aumento no volume de dados textuais, especialmente os não estruturados, gerados por pessoas que expressam opiniões através de várias plataformas de mídia social e *web,* por diferentes razões. Os autores afirmam que as montanhas desses dados textuais, inicialmente poderiam ser equiparadas a lixo - que precisaria ser descartado de tempos em tempos. No entanto, com o avanço na capacidade de armazenamento, acompanhado pela crescente sofisticação em ferramentas de mineração de dados, surgiram novas oportunidades e desafios para analisar e derivar *insights* úteis desses dados. Sharma, Nigam e Jain (2014) apontam também que, quando um grande número de avaliações está disponível para um único produto, fica difícil para o consumidor ler todas as resenhas e tomar uma decisão.

O crescimento das áreas de inteligência artificial e mineração de textos, porém, permitiu realizar, de forma automática, a identificação de comentários relevantes através de técnicas de análise de sentimentos e mineração de opinião. Segundo Sarakit et al. (2015), com o aumento do uso de mídias sociais, essa análise automática

passou a permitir que empresas capturem os sentimentos de seus clientes (ou potenciais clientes) sobre seus produtos e serviços, ao analisar as avaliações.

Para İşgüder-Şahin, Zafer e Adah (2014), a análise de sentimentos consiste em extrair o sentido, a polaridade e visão subjetiva do autor do texto não estruturado. E a detecção de polaridade é um tipo de análise que só está interessada em avaliar textos por meio de negatividade e positividade. Da mesma forma, Sharma, Nigam e Jain (2014) descrevem a mineração de opinião como uma tarefa responsável por identificar sentimentos expressos em comentários positivos ou negativos, analisando um grande número de documentos. Sendo assim, a principal tarefa da análise de sentimentos é a classificação de documentos e detecção de polaridade.

Apresentada essa discussão inicial, fica claro a existência de oportunidades para explorar comentários em sites e aplicativos, utilizando técnicas de mineração de texto. O objetivo geral do presente trabalho é propor um modelo capaz de identificar diferentes assuntos em dados textuais e classificá-los de acordo com a polaridade da opinião do autor - positiva ou negativa. Buscando testar esse modelo, são propostos dois métodos distintos para detectar assuntos e predizer a polaridade em textos: um método sem a aplicação de aprendizagem de máquina e outro com aplicação de aprendizagem de máquina.

Como objetivos específicos, este trabalho buscou:

A. Facilitar o entendimento de uma grande quantidade de informações textuais
B. Aplicar técnicas de aprendizagem de máquina para modelagem de tópicos e análise de sentimentos.
C. Identificar pontos positivos no modelo, investigando os resultados das diferentes técnicas aplicadas.

Com o modelo proposto é possível sintetizar um grande volume de informações textuais em um pequeno conjunto de tópicos, subdivididos entre polaridades positivas e negativas. Esta organização pode facilitar para empresas a extração de conhecimento sobre grandes conjuntos de dados e auxiliar a separação de textos filtrando por determinados assuntos e sentimentos, automaticamente.

As próximas sessões detalhem o modelo proposto. A Sessão 2 explica as técnicas de aprendizagem de máquina utilizadas. A Sessão 3 discute sobre trabalhos relacionados. A Sessão 4 explica o modelo proposto, enquanto a Sessão 5 relata sua

implementação. A Sessão 6 apresenta os resultados obtidos. E, por fim, a Sessão 7 apresenta a conclusão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

A análise de texto pode envolver a classificação do texto em algum tipo de categoria. Classificação de textos é o nome dado à tarefa de definir um rótulo ou categoria a um texto ou documento. Uma das tarefas mais comuns da classificação de textos é a análise de sentimentos, que nada mais é do que a detecção de um sentimento predominante no texto analisado. (JURAFSKY; MARTIN, 2018)

O objetivo da classificação é analisar uma única observação, extrair algumas características úteis e, assim, classificá-la em uma, dentre um conjunto de classes distintas. Um dos métodos utilizados para classificar textos é a aplicação de regras de escrita. Existem muitas áreas de processamento de linguagem natural em que sistemas, ou pelo menos partes deles, utilizam classificadores baseados em regras de escrita. (JURAFSKY; MARTIN, 2018)

Outra alternativa para a classificação são os algoritmos de aprendizagem de máquina supervisionados. Na aprendizagem supervisionada a entrada de dados do algoritmo é um conjunto de observações, onde cada observação está corretamente associada à uma saída (classe). O objetivo do algoritmo é aprender com os dados de entrada como mapear novas observações às saídas (classes) corretas. (JURAFSKY; MARTIN, 2018)

2.1 Modelagem de tópicos

2.1.1 LDA

Modelagem de tópicos é a tarefa da mineração de textos responsável por extrair assuntos dentro de uma coleção de documentos (corpus). O algoritmo mais comum em uso atualmente nas aplicações desta tarefa é o *Latent Dirichlet Allocation* (LDA) (BLEI E JORDAN, 2003). Lu et al. (2013) explicam que a ideia principal do algoritmo é que cada tópico seja modelado como uma distribuição de probabilidade sobre um conjunto de palavras, enquanto cada documento seja modelado como uma distribuição de probabilidade sobre um conjunto de tópicos.

Faleiros e Lopes (2016) analisam o LDA, explicando que as variáveis observáveis são os termos de cada documento, enquanto as variáveis não observáveis são as distribuições de cada tópico. A distribuição de Dirichlet é utilizada para amostrar as distribuições de tópicos. No processo generativo, o resultado da amostragem de Dirichlet é usado para alocar as palavras de diferentes tópicos que preencherão os documentos.

Faleiros e Lopes (2016) complementam que, desta forma, os tópicos são definidos como distribuições de probabilidades sobre um vocabulário fixo de palavras, enquanto os documentos são *bag-of-words*, surgindo da escolha aleatória das palavras pertencentes a uma distribuição de tópicos.

Para descrever o processo gerador do modelo LDA, Faleiros e Lopes (2016) descrevem os passos para a criação de um documento d_j da seguinte forma:

1. Amostre K multinomiais $\phi_k \sim Dir(\phi_k, \beta)$, um para cada tópico k.
2. Amostre *m* multinominais $\theta_j \sim Dir(\theta_j, \alpha)$, um para cada documento d_j.
3. Para cada documento d_j da coleção:
    a. Para cada palavra w_i do documento d_j:
        i. Associe um tópico para $z_{j,i}$ amostrado da distribuição de Dirichlet θ_j.
        ii. Amostre uma palavra w_i da distribuição $\phi_{z_{j,i}}$.

Segundo Faleiros e Lopes (2016), são utilizadas duas variáveis para representar as distribuições. A variável ϕ representa as palavras do vocabulário. A variável θ representa o número de tópicos. As duas variáveis são geradas pela distribuição de Dirichlet (Dir) com seus respectivos hiper-parâmetros β e α. Faleiros e Lopes (2016) explicam que, definindo um alto valor para α, cada documento conterá uma maior mistura de tópicos. Enquanto isso, um valor baixo para α indica uma mistura de poucos tópicos no documento. Da mesma forma, um valor alto para β significa que cada tópico poderá conter misturas de muitas palavras, enquanto um valor baixo para β indica que o tópico será formado por poucas palavras.

Faleiros e Lopes (2016) afirmam que, ao analisar as variáveis observáveis e não observáveis, o objetivo passa a ser descobrir as atribuições de tópicos para os documentos e as distribuições de documentos por tópicos e tópicos por termos. Sendo assim, o grande problema computacional do LDA é inferir $p(z, \phi, \theta | w, \alpha, \beta)$, onde w são todas as palavras observadas na coleção de documentos.

Para Faleiros e Lopes (2016), é possível resolver o problema computacional central inferindo a probabilidade a *posteriori* de todo o modelo. “[...] Esse cálculo de inferência pode ser feito pela soma da distribuição conjunta de todos os valores possíveis atribuídos às variáveis não observadas (todas as palavras da coleção)” (FALEIROS; LOPES, 2016). Porém, os autores explicam que o número de atribuições possíveis é exponencialmente grande, tornando esse cálculo intratável computacionalmente. Como forma de solução, existem vários métodos para aproximar a distribuição a *posteriori*. Entre os mais utilizados na literatura, estão o Amostrador de Gibbs e a Inferência Variacional.

2.2 Classificadores

2.2.1 Support Vector Machine

Support Vector Machine (SVM) é um método de classificação estatístico elaborado por Vladimir N. Vapnik, publicado nos anos 90. A ideia por trás do algoritmo é encontrar a melhor linha divisória que separe elementos de um conjunto pela sua classe. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

A linha divisória é chamada de hiperplano e o espaço entre o hiperplano e os elementos mais próximos é conhecido como “margem”. A margem é idealizada como um espaço vazio, sem nenhum elemento. O melhor hiperplano será aquele que maximize a largura da margem. Os elementos que estiverem sobre a fronteira da margem serão chamados de vetores de suporte. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Embora a ideia seja simples, sua implementação é complexa e envolve diversas áreas do cálculo, como análise vetorial, geometria analítica, multiplicadores de Lagrange e condições de Karush-Kuhn-Tucker. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Um dos grandes diferenciais do método SVM é a aplicação de *kernels*. Didaticamente, o SVM é apresentado solucionando problemas que podem ser separados por uma linha reta. Chamamos estes problemas de lineares. Porém, soluções lineares não são aplicáveis para conjuntos com mais de duas dimensões ou em que as informações estão muito misturadas. Para esse tipo de situação, o SVM

utiliza uma função de mapeamento chamada *kernel* - responsável por adaptar os elementos em um novo plano. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

O *kernel* é uma função matemática que aplica transformações nos vetores de entrada, de forma a separar melhor as instâncias a ponto de conseguir classificá-las. Não existe nenhuma dependência entre o SVM e a função *kernel* e, por isso, existem diversas opções de *kernel* a serem utilizados, como o linear, o polinomial e o *radial basis function*. (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

2.2.2 Naive Bayes

Segundo Nigam et al. (1998), Naïve Bayes é um algoritmo muito popular para classificação de textos. O algoritmo pertence ao grupo dos classificadores estatísticos que utilizam as frequências das palavras como atributos. O algoritmo é baseado na teoria de Bayes e é chamado de *naive* (ingênuo), pois assume algumas suposições sobre como os atributos se relacionam. As principais suposições são que todas as probabilidades dos atributos de entrada são independentes e a sua ordem não importa.

Segundo Pang, Lee e Vaithyanathan (2002), uma abordagem para classificação de textos é atribuir a um determinado documento *d* a classe *c**, calculado através da Equação 1, onde $P(c|d)$ é resolvida através da regra de Bayes, dada pela Equação 2.

Equação 1

$$c^* = argmax_c\, P(c|d)$$

Equação 2

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Jurafsky e Martin (2018) explicam que, através da decomposição destas equações, aplicando as tais suposições, obtém-se a equação final do Naive Bayes, dada pela Equação 3, onde C representa o conjunto de classes, e F, as frequências dos atributos de entrada.

Equação 3

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c) \prod_{f \in F} P(f|c)$$

Uma explicação livre seria que o Naïve Bayes estima a probabilidade de um documento pertencer a todas as classes, analisando as palavras de entrada, e define como sua classe a que atingiu a maior probabilidade.

Pang, Lee e Vaithyanathan (2002) analisam que, apesar da sua simplicidade e do fato de que a suposição de independência condicional claramente não se sustenta em situações do mundo real, a categorização de textos baseada em Naive Bayes ainda tende a apresentar um bom desempenho.

3 TRABALHOS RELACIONADOS

As avaliações de consumidores sobre produtos e serviços adquiridos se tornaram informações valiosas nos dias de hoje. Prestar atenção aos comentários de clientes sobre seus produtos e serviços é uma questão de sobrevivência para qualquer empresa. Consumidores compartilham suas experiências através de avaliações *online* que ficam disponíveis a outros consumidores. Assim, novos clientes podem levá-las em consideração na hora de optar por fazer negócios ou não – o que torna a interação *online* um fator muito importante para a conversão de clientes. Por isso as empresas devem estar muito atentas às opiniões recebidas, lendo e respondendo adequadamente, de forma a criar um ambiente confiável para o consumidor. Mas quando o volume de informações é enorme, esta tarefa pode se tornar muito difícil. Na literatura, a mineração de textos vem sendo muito explorada para contornar este problema.

Rane e Kumar (2018) analisam a interação entre clientes e empresas de companhias aéreas americanas, propondo classificação automática de *tweets* como forma de substituir pesquisas de satisfação de clientes. O artigo disserta sobre a aplicação de diferentes algoritmos de aprendizagem de máquina para determinar a polaridade dos comentários. E conclui que a acurácia obtida nos classificadores é alta o suficiente para ser empregada como substituta das pesquisas de satisfação.

Utilizando comentários textuais de avaliações de hotéis para aplicar análise de sentimentos e mineração de opinião, Zvarevashe e Olugbara (2018) propõem um modelo em que os comentários são rotulados com base numa pontuação determinada por um algoritmo de detecção de polaridade de sentimento. Foram utilizados quatro algoritmos de classificação para treinar e testar o conjunto de dados, dentre os quais, Naïve Bayes Multinomial obteve a maior precisão. O trabalho afirma que a análise de

sentimento proposta pode ser incorporada em um sistema de tecnologia de hotéis, e que pode ajudar a melhorar o gerenciamento do relacionamento com o cliente.

Já Salinca (2015), estudou a análise de sentimentos sobre empresas utilizando comentários do *Yelp*. O autor apresenta os resultados de vários algoritmos de aprendizagem de máquina para classificar avaliações, aplicando análise de sentimentos e técnicas de processamento de linguagem natural. O artigo também testa uma segunda abordagem utilizando um dicionário personalizado e detecção de polaridade com léxicos de sentimentos. Neste segundo procedimento, a polaridade de cada palavra em cada sentença foi dada pelo SentiWordNet. As somas dos sentimentos de todas as palavras da sentença definem a sua polaridade. Salinca (2015) conclui que a melhor classificação foi obtida com a primeira abordagem, utilizando *Linear SVC* e *SGD*. A abordagem com dicionários não obteve bons resultados e não foi considerada útil.

Ao fazer uso de dicionários de palavras para extrair conceitos relevantes para termos específicos de domínios, Kim e Cavedon (2011) propõem um modelador de tópicos baseado na associação entre tópicos e dicionários contendo termos relevantes para cada conceito de domínio.

Tang et al. (2016) sugerem um modelo baseado em LDA capaz de extrair tópicos relevantes de posts realizados por estudantes no site da Universidade BBS, na China. O estudo foca no ganho de qualidade para extração de tópicos através da identificação de termos relevantes que são definidos a cada tópico, utilizando estas informações para gerar um modelo SVM capaz de identificar tópicos com maior precisão.

Analisar textos extraídos do Twitter, a respeito do atendimento da população na cidade de Surabaya, é o modelo apresentado por Azis et al. (2018). O processo de classificação de sentimento compara a análise léxica, através do SentiWordNet, com aprendizado de máquina, utilizando SVM. Após detectar polaridades, o algoritmo LDA é utilizado para extração de tópicos em cada grupo de sentimentos. O trabalho aponta que a utilização de aprendizagem de máquina para análise de sentimentos produz melhores resultados que a análise léxica.

De forma similar, Alamsyah et al. (2018) coletaram, no Twitter, dados relacionados à empresa Uber para minerar opiniões. O modelo proposto utiliza Naive Bayes para análise de sentimentos. Os comentários são separados em dois arquivos, de acordo com as polaridades detectadas, e depois são extraídos tópicos relevantes

com LDA em cada arquivo. Ao final, o trabalho analisa os termos dos tópicos extraídos para sentimentos positivos e negativos. Conclui que obteve bons resultados mas questiona a dificuldade de gerar métricas de precisão.

Os trabalhos de Rane e Kumar (2018) e Zvarevashe e Olugbara (2018) demonstram que é possível analisar sentimentos com aprendizagem de máquina, e são valiosos por indicar algoritmos eficientes para estes processamentos.

O modelo proposto pelo presente trabalho é similar aos de Alamsyah et al. (2018) e Azis et al. (2018. Nestes últimos, porém, o modelo de extração de tópicos é aplicado em textos separados de acordo com o sentimento. Há um modelo extrator de tópicos para textos classificados como positivos e outro para negativos. Já este artigo propõe o inverso: identificar tópicos relevantes em textos e detectar o sentimento em relação ao tópico identificado.

Como forma de comparar o modelo proposto com aprendizagem de máquina, este trabalho aplica outro método mais simples, inspirado em Kim e Cavedon (2011). Através de dicionários de palavras associados a tópicos, os assuntos são extraídos pela frequência de termos encontrados. Para a análise de sentimentos neste modelo simples, também são utilizados léxicos de sentimentos, similar aos propostos por Salinca (2015) e Azis et al. (2018). O Quadro 1 apresenta um comparativo entre os trabalhos relacionados.

Quadro 1 – Comparativo entre trabalhos relacionados

Trabalho	**Proposta**	**Base de dados**	**Extração de Tópicos**	**Analise de sentimentos**
Rane e Kumar (2018)	Analisar sentimentos em tweets de companhias aéreas	Twitter US Airline Sentiment		Decision Tree, Random Forest, Logistic Regression, SVM, Naive Bayes, AdaBoost e KNN
Zvarevashe e Olugbara (2018)	Utilizar avaliações de hotéis para aplicar análise de sentimentos com a mineração de opinião	OpinRank		Naïve Bayes, SVM e Composite Hypercubes on Iterated Random Projections (CHIRP)

Salinca (2015)	Utilizar algoritmos de aprendizado de máquina para classificar as avaliações do Yelp, aplicando análise de sentimentos e técnicas de processamento de linguagem natural	Yelp Challenge Dataset		Naïve Bayes, SVM e analise léxica com SentiWordNet
Kim e Cavedon (2011)	Utilizar dicionários de palavras para extrair conceitos relevantes para termos específicos de domínios		Dicionários de palavras com bag-of words e WordNet	
Tang et al. (2016)	Extrair tópicos relevantes de posts realizados por estudantes no site da Universidade BBS	BYR BBS	LDA e SVM	
Azis et al. (2018)	Analisar textos extraídos do Twitter sobre o atendimento da população na cidade de Surabaya	Twitter	LDA	SVM, análise léxica com SentiWordNet
Alamsyah et al. (2018)	Analisar dados relacionados ao Uber no Twitter para minerar opiniões	Twitter	LDA	Naive Bayes
Trabalho proposto	Identificar assuntos em comentários de restaurantes e avaliar a polaridade sobre cada assunto identificado	TripAdvisor	LDA e dicionários de palavras	Naive Bayes, SVM e analise léxica com dicionários SentiLex e Oplexicon

Fonte: Elaborado pelo autor

4 MODELO PROPOSTO

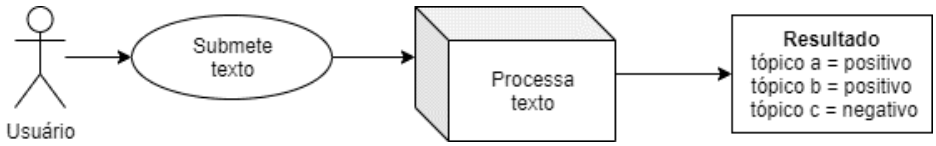
4.1 OTTO: Modelo de extração de tópicos e sentimentos

Com o avanço da internet e das redes sociais, a forma como interagimos com empresas mudou consideravelmente nos últimos anos. Milhares de dados textuais são produzidos a todo momento. Tais dados estão, muitas vezes, na forma de comentários, expressando opiniões de consumidores, trazendo informações que podem ser cruciais para companhias. Porém, devido ao alto volume, é muito difícil para uma empresa dar atenção a todos os dados recebidos.

Com o objetivo de atender a este problema, é proposto um modelo capaz de identificar assuntos relevantes e o respectivo sentimento do autor (positivo ou negativo) em textos. Ao invés de dizer o assunto e sentimento geral do texto, o modelo

visa detectar os principais tópicos tratados, descrevendo o sentimento do autor para cada assunto encontrado.

Figura 1 – Modelo do Trabalho Proposto



Fonte: Desenvolvido pelo autor.

Na Figura 1 o módulo “Processa texto” representa o modelo proposto. Para a execução deste módulo são necessários métodos capazes de detectar tópicos e analisar sentimentos. A Figura 2 retrata o processamento interno do módulo.

Figura 2 – Módulo de Processamento de Textos

Fonte: Desenvolvido pelo autor.

De modo geral, um novo texto entra no módulo “Processa texto” e é submetido a dois processos: no primeiro, o texto é analisado com o extrator de tópicos, responsável por identificar o assunto predominante; já no segundo, o texto é processado no modelo de análise de sentimentos, para identificar a polaridade. Ao final, o módulo resulta no assunto identificado e na polaridade do texto.

Para fins comparativos, foram abordados dois métodos distintos para elaborar o módulo. O primeiro método faz uso de dicionários de palavras para encontrar tópicos e léxicos de sentimentos para identificar polaridades. O segundo método aplica técnicas de aprendizagem de máquina, nas quais modelos são treinados para realizar inferências.

Nas próximas seções serão explicadas cada uma das abordagens.

4.2 Método 1: Sem aprendizagem de máquina

O objetivo deste método é elaborar um procedimento capaz de identificar assuntos e sentimentos de um texto através de dicionários de palavras.

4.2.1 Extração de tópicos

Foi necessário definir previamente os tópicos relevantes a serem identificados. Cada tópico foi associado a um dicionário contendo palavras que caracterizam o assunto. Este método busca identificar os assuntos através da frequência de palavras do texto nos dicionários. O dicionário com a maior quantidade de palavras do texto de entrada é o tópico principal do texto.

4.2.2 Análise de sentimentos

Um léxico de sentimentos de palavras foi utilizado para identificar a polaridade de cada termo presente no texto. O léxico de sentimentos resulta no valor +1 para palavras positivas e -1 para negativas. A polaridade do texto é calculada somando as pontuações obtidas, e se o resultado for maior que 0, o texto é considerado positivo. Caso contrário, é negativo.

4.3 Método 2: Com aprendizagem de máquina

4.3.1 Extração de tópicos

Este método aplicou a extração de tópicos e a análise de sentimentos com algoritmos de aprendizagem de máquina. Para a extração de tópicos foi utilizado *Latent Dirichlet Allocation* (LDA), cujo o modelo gerado é capaz de identificar a distribuição de probabilidade para tópicos em um documento. O tópico com a probabilidade mais significativa é escolhido para representar o texto.

4.3.2 Análise de sentimentos

Para a análise de sentimentos foram utilizados modelos treinados com os algoritmos Naive Bayes e SVM. Para identificar a polaridade, o texto é submetido a

um pré-processamento antes de ser enviado ao analisador de sentimentos. O resultado são as probabilidades de o texto de entrada pertencer a cada sentimento (positivo ou negativo). O sentimento que apresentar a maior probabilidade é escolhido para representar a polaridade do texto.

4.4 O Modelo preditivo

Os dois métodos de extração e análise de sentimentos foram testados no módulo “Processa texto” (Figura 1). O texto de entrada pode ter mais de um assunto e apresentar sentimentos diferentes para cada um deles. Para identificar os diferentes assuntos, o texto de entrada é decomposto em sentenças. Cada sentença é submetida ao módulo “Processa texto” e, dessa forma, atrelada a apenas um assunto e uma polaridade.

5 IMPLEMENTAÇÃO

A grande maioria dos trabalhos relacionados analisa textos na língua inglesa. Fazendo uma simples comparação entre o inglês e português, encontramos diferenças que podem dar alguma vantagem à língua inglesa na mineração de textos, pois não utiliza acentos, enquanto no português um simples acento pode mudar totalmente o sentido da palavra. As conjugações de verbos em inglês são mais simples e geralmente acompanhadas por auxiliares padrões para apontar o tempo verbal. No português, cada tempo verbal possui uma conjugação específica através de sufixos, que mudam para cada pronome.

A mineração de textos para o inglês também está mais madura em relação a ferramentas e *frameworks*. Esta maturidade facilita a obtenção de ferramentas que auxiliam a análise, como listas de *stop-words*, dicionários, léxicos de sentimentos, analisadores de sintaxe e *part-of-speech taggers*. Em português, as opções são limitadas e algumas ferramentas utilizam as próprias versões em inglês, mas aplicando a tradução.

Com base nisso, a mineração de textos escritos em português é desafiadora e se torna um objeto de pesquisa válido. Com o intuito de aplicar as técnicas de extração de tópicos e análise de sentimentos, se propõe a implementação de um modelo preditivo que analisa pequenos textos.

Modelo esse que foi desenvolvido e implantado em um sistema *web*[1] onde usuários submetem comentários e o sistema extrai assuntos relevantes com suas respectivas polaridades de opinião. Para quantificar a qualidade do modelo, ao final do processo o site questiona ao usuário se as predições estão corretas. Os resultados foram quantificados e serão discutidos no decorrer deste artigo.

5.1 Base de dados

Para modelar os algoritmos de predição foi necessária uma base de dados com muitas informações. Para obter uma grande massa de dados se propõe a extração de comentários sobre restaurantes do site TripAdvisor. A escolha é justificada pelo fato de o site ser uma referência em aplicativos de avaliações turísticas, onde qualquer pessoa pode comentar sobre hotéis, restaurantes e locais. Além disso, é utilizado globalmente, contando com milhares de usuários em todas as regiões do planeta.

A mineração de textos sugere que as bases de dados utilizadas para treinamento contenham muitos registros. Por ser necessário um grande número de avaliações, foram filtrados restaurantes de alguns dos principais destinos turísticos do Brasil. As cidades selecionadas estão entre as dez mais visitadas do país no ranking do TripAdvisor. São elas: Rio de Janeiro, Salvador, Porto de Galinhas, Porto Seguro, Maceió e Porto Alegre.

Ao total, foram capturadas 112.759 avaliações em Porto Alegre, 67.050 em Porto Seguro, 384.336 no Rio de Janeiro, 39.696 em Porto de Galinhas, 84.104 em Maceió e 99.818 em Salvador. A base unificada com todas as avaliações conta com 776.185 comentários.

Por pretender aplicar técnicas de aprendizagem de máquina supervisionada para análise de sentimentos, uma nova base de dados teve que ser elaborada. Esta base necessitava de instâncias rotuladas pelo sentimento do autor expressos como positivos e negativos. Utilizando os textos e notas obtidos com o *crawler*, foram extraídas sentenças e agrupadas em uma nova base. Destas sentenças, foram etiquetadas manualmente 103.912 instâncias, sendo 62.604 positivas e 41.308 negativas.

[1] Publicado em http://tcc.e-mine.com.br. Acesso em 22/11/2019.

5.2 Crawler

Para extrair as avaliações do TripAdvisor foi desenvolvido um *crawler* capaz de navegar pelo site, realizando buscas por restaurantes para cada cidade selecionada. O *crawler* capturou comentários e notas sobre cada estabelecimento encontrado. As avaliações obtidas foram agrupadas em uma base no formato JSON, onde, posteriormente, foram aplicados os métodos propostos. Foram implementados *crawlers* para cada uma das cidades selecionadas.

5.3 Método 1 – Sem aprendizagem de máquina

5.3.1 Extração de tópicos

O método de extração de tópicos utilizando dicionários consistiu em buscar cada palavra presente no texto dentro dos dicionários de tópicos. O tópico no qual o dicionário possui a maior quantidade de palavras localizadas é o representante do texto.

O método foi implementado em Python utilizando o WordNet para resolver ambiguidades, entre outras tarefas. Quatro tópicos foram definidos, de forma manual. Para cada tópico, um dicionário foi criado contendo palavras que o identificassem, com base no conhecimento de domínio do autor.

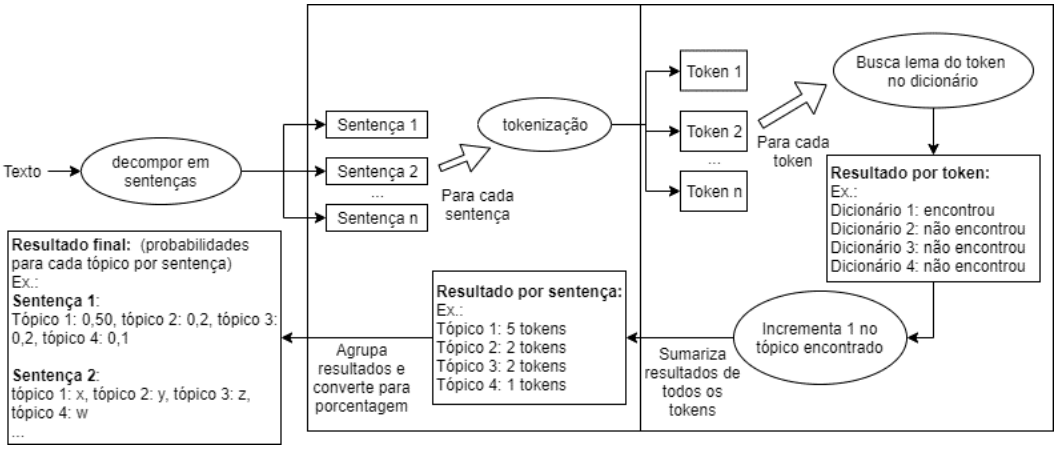
Alguns problemas precisaram ser contornados. É possível uma palavra ser examinada no dicionário e não ser encontrada, mas algum sinônimo pode estar presente. Outro problema são as dificuldades em encontrar palavras no plural ou verbos conjugados. Além disso, o suporte para português do WordNet só reconhece lemas de palavras.

Como forma de contornar o problema, durante a fase de pré-processamento foram removidas *stop-words*, pontuações e palavras estrangeiras, além de lematizar todos os termos. Para realizar o processo de lematização, foram utilizadas as bibliotecas *Spacy* integradas com a *nlp-stanford*. *Spacy* é uma biblioteca muito versátil para NLP, capaz de executar diversas tarefas de pré-processamento com poucos comandos.

Com o auxílio do *WordNet*, os dicionários foram modificados para incluir sinônimos e antônimos de palavras já presentes. Isto foi realizado na tentativa de

aumentar a dimensão de cada dicionário. Através de funções do *WordNet* também foi possível decompor o texto em suas sentenças. A biblioteca *Spacy* tokeniza a sentença, convertendo cada palavra em um token já com o lema e o POS Tag correspondente. Cada token é submetido ao WordNet através do lema, que por sua vez, devolve sinônimos e antônimos correspondentes. Todos os termos obtidos são examinados nos dicionários de cada tópico. Quando um termo é encontrado em algum dicionário, o tópico correspondente ao dicionário recebe um ponto (incremento de 1). O tópico que obtiver mais pontos é escolhido para representar aquela sentença. A Figura 3 descreve o fluxo do texto submetido no modelo proposto.

Figura 3 – Implementação do Extrator de Tópicos

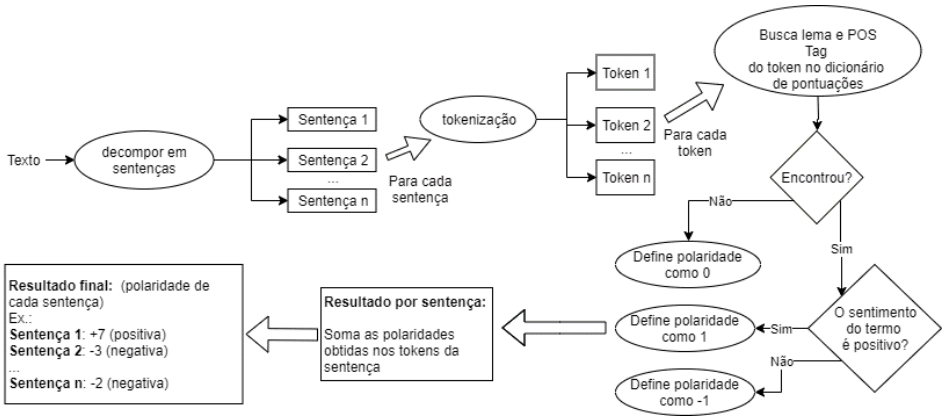


Fonte: Elaborado pelo autor

5.3.2 Analise de sentimentos

Para descobrir as polaridades dos textos foram utilizados os léxicos de sentimentos SentiLex e Oplexicon. De forma similar ao que foi aplicado na extração de tópicos, os lemas também foram utilizados na análise de sentimentos. Porém, para resolver ambiguidades na busca por termos no léxico de sentimentos, é necessário identificar o tipo sintáxico de cada palavra, ou seja, se o termo é substantivo, verbo ou adjetivo. Cada palavra é submetida ao léxico de sentimentos através do lema e do tipo sintáxico. Encontrados os termos correspondentes, as pontuações retornam como +1 para termos com sentimentos positivos e -1 para negativos. Caso o termo não seja encontrado, o valor 0 retorna. Ao final, as pontuações são somadas e o sinal do resultado define a polaridade da sentença. A Figura 4, descreve o fluxo do texto de entrada no modelo proposto.

Figura 4 – Implementação da Análise de Sentimentos



Fonte: Elaborado pelo autor

5.4 Método 2 – Com aprendizagem de máquina

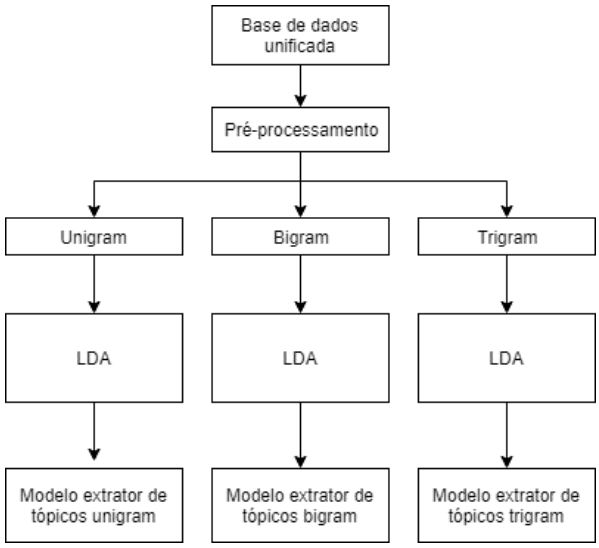
5.4.1 Extração de tópicos

A abordagem com aprendizagem de máquina foi desenvolvida em Python, onde foram empregadas as bibliotecas NLTK, Sklearn e Gensim. NLTK é uma plataforma para processamento de linguagem natural muito versátil, que conta com recursos para tokenização, lematização, stemming, classificação, entre outras funcionalidades. A biblioteca Gensim possui uma implementação do LDA e conta com recursos auxiliares que facilitam a interpretação dos resultados da extração de tópicos. Também foi utilizado o algoritmo LDA da biblioteca MALLET, para averiguar uma possível melhora na qualidade do modelo. A biblioteca MALLET emprega o Amostrador de Gibbs para elaborar as distribuições Dirichlet, enquanto Gensim utiliza Inferência Variacional.

Para a modelagem da extração de tópicos com LDA todos os textos capturados com o *crawler* são unificados em um único *dataset*. Na etapa de pré-processamento o *dataset* sofre tokenização, são removidas pontuações, *stop-words* e palavras que não sejam substantivos ou verbos, e, por fim, aplica-se *stemming* nas restantes.

O modelo LDA foi gerado a partir do treinamento da base de dados com unigram, porém, testes foram feitos aplicando bigram e trigram. A Figura 5 descreve o processo de treinamento do modelo LDA.

Figura 5 – Treinamento do modelo de extração de tópicos com LDA

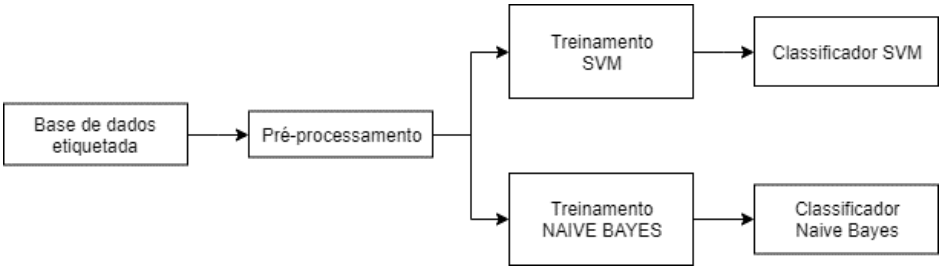


Fonte: elaborado pelo autor.

5.4.2 Análise de sentimentos

Para a análise de sentimentos foram empregados algoritmos da biblioteca Sklearn. Os modelos foram gerados a partir da base de sentenças classificadas manualmente, descrita na sessão 5.1. O pré-processamento eliminou pontuações, *stop-words* e aplicou *stemming* e TF-IDF. Seguindo o que foi pontuado por Rane e Kumar (2018), palavras como "não" e “nenhum” não foram removidas dos textos, pois adicionam um significado à frase. A base resultante foi submetida ao treinamento dos classificados Naive Bayes e Linear SVM, ambos implementados na biblioteca Sklearn. O modelo Linear SVM atingiu acurácia de 99%, mas não manteve a mesma qualidade quando testado com dados novos. O modelo Naive Bayes obteve acurácia de 95,23%, e foi o escolhido para ser utilizado no projeto. A Figura 6 descreve o processo de treinamento da análise de sentimentos.

Figura 6 – Treinamento do Modelo de extração de tópicos com LDA



Fonte: elaborado pelo autor.

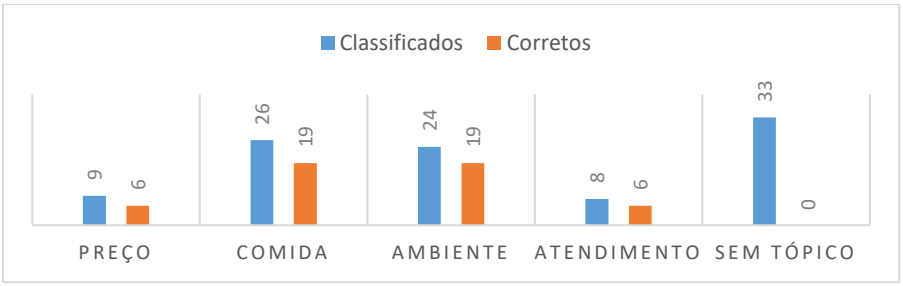
6 TESTES E RESULTADOS

6.1 Abordagem de extração de informações sem aprendizagem de máquina

O método baseado em dicionário de palavras apresentou resultados insatisfatórios. Muito tempo foi gasto com a lematização de palavras, uma tarefa aparentemente trivial, mas muito importante para possibilitar buscas nos dicionários e no WordNet. A lematização do WordNet acabou apresentando resultados irregulares, não identificando verbos conjugados, por exemplo. A lematização do Spacy, outra ferramenta muito popular em NLP com Python, apresentou resultados melhores, porém muitas vezes resultando em palavras que não existiam. Porém, Spacy possibilita integrar a biblioteca *nlp-stanford*, que, por sua vez apresentou ótimos resultados. O grande problema acabou sendo o tempo consumido para executar os processos. Uma máquina com processador i7, 8GB de RAM, não obteve sucesso ao processar os 700 mil textos.

Os testes foram executados com limite de 100 sentenças. Dentre estes, 50 foram identificados no tópico correto. Trinta e três sentenças não foram identificadas em nenhum tópico. Frequentemente foram encontradas palavras importantes para identificar um tópico, mas que não estavam presentes no dicionário. O trabalho manual de acrescentar novas palavras aos dicionários de tópicos acabou sendo repetitivo. No fim, os dicionários de tópicos acabaram desbalanceados, e a maioria dos comentários era atribuída a um determinado tópico, ignorando outros que fossem mais propícios. O Gráfico 1 apresenta o resultado da extração de tópicos.

Gráfico 1 – Sentenças classificas por tópico



Fonte: Elaborado pelo autor.

A análise de sentimentos também apresentou resultados tendenciosos com léxicos de sentimentos. Os léxicos de sentimentos possuem valores pré-definidos

para cada palavra, sendo que muitas palavras comuns no contexto analisado não constavam nos arquivos ou possuíam uma atribuição que não correspondia ao significado do termo. Das 100 sentenças, 51 não tiveram qualquer palavra reconhecida nos léxicos. Como nesses comentários nenhuma palavra foi encontrada, as respectivas polaridades acabaram sendo 0 (neutras). Dentre as 49 sentenças reconhecidas, 7 foram classificadas erradas. A Tabela 1 apresenta os resultados em detalhes.

Tabela 1 – Resultados da classificação

	Classificados corretamente	**Classificados com erro**	**Não classificou**
Extração de tópicos	50	17	33
Análise de sentimentos	42	7	51

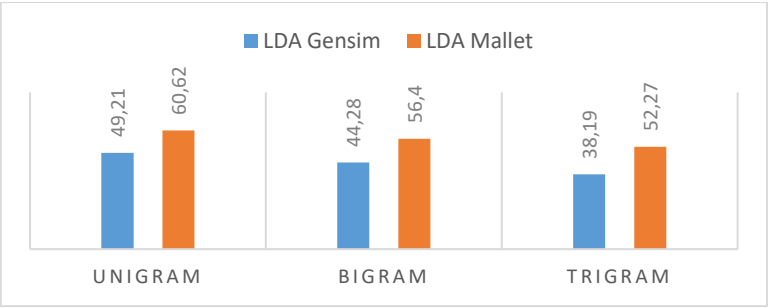
Fonte: Elaborado pelo autor.

6.2 Extração de tópicos com aprendizagem de máquina

O melhor resultado do modelo LDA foi obtido mantendo no *dataset* apenas substantivos ou verbos e aplicando *stemming* nas palavras que permaneceram. O número de tópicos definido como parâmetro para o algoritmo LDA foi de 10 tópicos. O valor foi escolhido empiricamente, avaliando as mudanças na coerência ao aumentar a quantidade de tópicos gradativamente.

A versão do algoritmo LDA da biblioteca MALLET gerou modelos melhores que a biblioteca Gensim. O modelo unigram gerado obteve coerência de 60%, enquanto o modelo com Gensim obteve 49%. O Gráfico 2 ilustra a diferença de qualidade entre os modelos gerados pelas duas bibliotecas.

Gráfico 2 – Valores de coerência para os modelos gerados



Fonte: Elaborado pelo autor.

O grande problema encontrado na abordagem de extração de tópicos foi a dificuldade em compreender e dar títulos ou significados coerentes para cada tópico obtido. Este processo é realizado por quem analisa os resultados, de forma manual, com base no conhecimento do domínio dos textos e nas distribuições de probabilidades resultantes.

6.3 Análise de sentimentos com aprendizagem de máquina

O modelo gerado com *Linear SVM* obteve acurácia de 99,22%. Um número tão alto geralmente indica *overfitting*. Para verificar, o modelo foi testado com uma nova base de dados com 100 novas instâncias, classificadas manualmente, das quais acertou apenas 55%. O modelo Naive Bayes obteve acurácia de 95,23%, mas apresentou resultados melhores na base de testes (91%). A Tabela 2 detalha os resultados.

Tabela 2 – Métricas de avaliações dos modelos Naive Bayes e SVM

	Naive bayes			**SVM**		
	precisão	**recall**	**f1-score**	**precisão**	**recall**	**f1-score**
positivo	0,96	0,97	0,97	0,99	0,98	0,99
negativo	0,95	0,94	0,95	0,97	0,99	0,98

Fonte: Elaborado pelo autor.

6.4 Testes com usuários

O sistema desenvolvido utilizou o modelo LDA unigram gerado com a biblioteca MALLET para extrair tópicos e o classificador Naive Bayes para analisar sentimentos. O sistema permitiu a um usuário submeter um comentário através de um site. O comentário era processado e os resultados avaliados pelos usuários através de questionário.

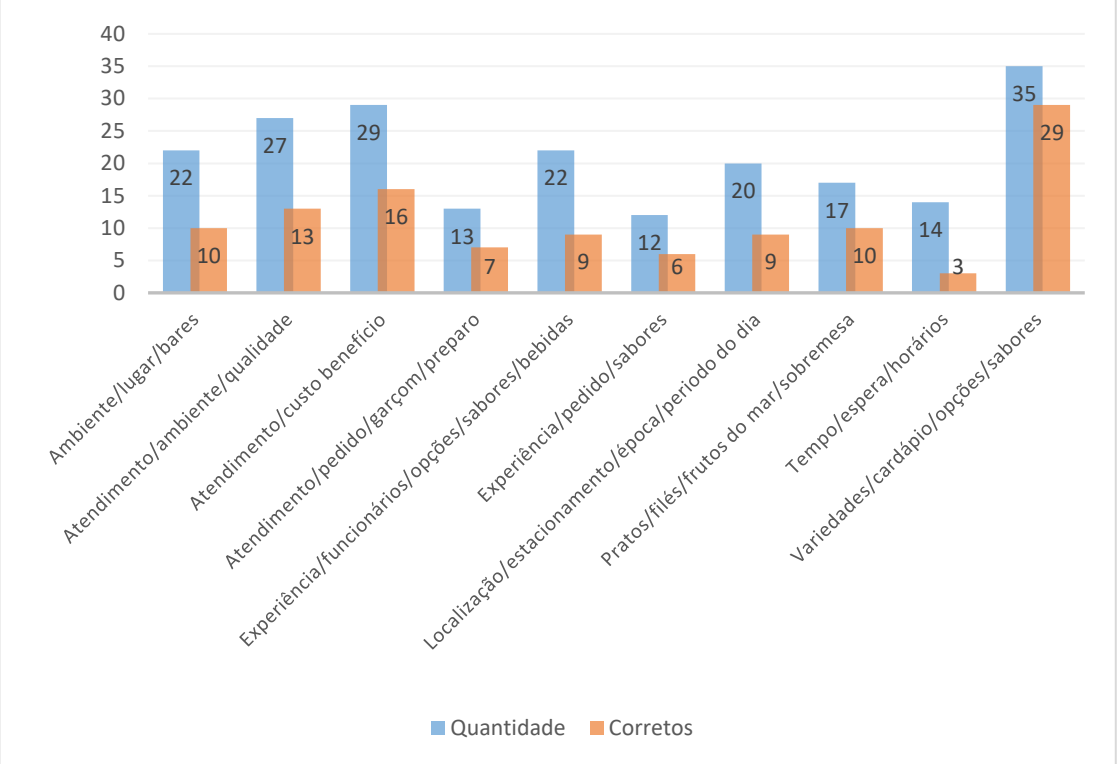
A primeira etapa do questionário foi perguntar ao usuário o assunto que as sentenças analisadas tratavam. Como o resultado do LDA é uma distribuição de probabilidades para cada tópico, o sistema elege como o assunto principal do texto o tópico que apresenta a maior probabilidade.

Porém, para validar os resultados, o usuário pôde escolher o assunto tratado, para, posteriormente se comparar os resultados enviados com os tópicos elegidos pelo sistema. Para isso, o usuário era induzido a escolher entre os três tópicos com

maior probabilidade resultantes da distribuição, com uma opção adicional de “nenhum”, caso não houvesse correspondência ao assunto.

A pesquisa contou com 75 avaliações diferentes, totalizando 211 sentenças analisadas. De acordo com as respostas dos entrevistados, dessas 211 sentenças, presumindo que o assunto principal é o tópico com maior probabilidade encontrado, o sistema foi capaz de acertar 112 tópicos. Os resultados são ilustrados no Gráfico 3.

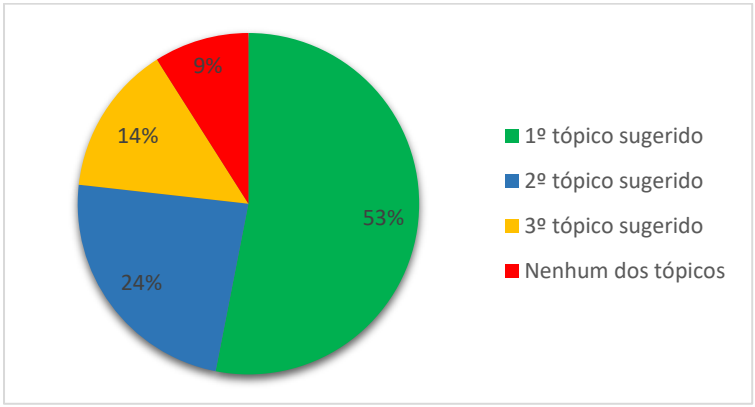
Gráfico 3 – Resultado da extração de tópicos



Fonte: Elaborado pelo autor.

Analisando os 99 tópicos restantes, os entrevistados apontaram que em 50 casos o assunto correto era o tópico com a 2ª maior probabilidade. E entre os 49 casos restantes, em 30 casos o assunto correto era o tópico com a 3ª maior probabilidade. Dos 213 textos, 19 sentenças não foram identificadas entre os 3 tópicos com probabilidades mais significativas.

Gráfico 4 – Classificações Corretas x Tópico sugerido



Fonte: Elaborado pelo autor.

O Gráfico 4 ilustra a distribuição de classificação correta de tópicos, onde os tópicos sugeridos representam os 3 tópicos com maiores probabilidades de representar o assunto do texto. Estes resultados são interessantes, pois demonstram que existe uma coerência na extração de tópicos. O sistema conseguiu descobrir que o texto pertencia a algum dos 3 tópicos sugeridos em 91% dos casos. O problema pode estar na nomenclatura do tópico, que foi definida manualmente pelo autor. Observando algumas respostas, notou-se que alguns entrevistados não conseguiram entender o sentido de alguns tópicos. Por exemplo, na sentença “primeira vez na casa e fomos excelentemente atendidos pelo Mael” o sistema apontou como pertencendo ao tópico “Experiência/funcionários/bebidas/ opções/sabores”. Mas para o usuário o tópico que melhor descreve a frase era “Atendimento/ambiente/qualidade”. Analisando a frase, o comentário fala sobre a sobre o excelente atendimento dado pelo funcionário Mael, o que demonstra que a atribuição ao tópico não estava errada.

Quanto à análise de sentimentos, os entrevistados apontaram que o sistema acertou 173 sentenças. Com isso, a análise de sentimentos proposta atingiu uma acurácia de 81,99%. Porém, analisando as classificações, pode-se observar que diversas instâncias negativas foram classificadas erradas. Os testes finais apresentaram uma precisão de 62,79% para a classe negativa, o que significa que, de todos as sentenças negativas classificadas, apenas 62,79% realmente eram negativas. Já o recall da classe negativa ficou em 55,1%, indicando que de todas as instâncias negativas existentes, apenas 55,1% foram classificadas corretamente. Com esses números já é possível entender que o modelo de análise de sentimentos não

foi bom para identificar sentenças negativas. A Tabela 3 detalha os resultados do modelo.

Tabela 3 – Precisão e recall do modelo Naive Bayes

	precisão	**Recall**
positivo	0,8690	0,9012
negativo	0,6279	0,5510

Fonte: Elaborado pelo autor

7 CONCLUSÃO

A proposta deste trabalho foi verificar a viabilidade da aplicação de técnicas de mineração de texto na língua portuguesa, agregando à pesquisa na área. As dificuldades para a análise de textos em português são diversas. Além da própria língua possuir características que dificultam a análise, as ferramentas para auxiliar a mineração de texto em português ainda não estão maduras. A maioria dos trabalhos relacionados analisam textos na língua inglesa, com o auxílio de diversos *frameworks* específicos para a língua. Era esperado com este trabalho apresentar uma visão melhor deste cenário, obtendo sucesso ao utilizar as ferramentas e técnicas do estado da arte.

O método proposto, aplicando aprendizagem de máquinas, apresentou resultados próximos aos observados nos artigos relacionados. Embora as métricas tenham sido altas nos modelos de análise de sentimentos, quando testados com usuários reais, não mantiveram uma ótima performance.

É importante salientar a importância de identificar textos negativos. É provável que empresas se interessem mais por compreender as opiniões negativas, de forma a conseguir agir rapidamente para modifica-las. O modelo Naive Bayes apresentou uma alta tendência em classificar instâncias como positivas. Como a grande maioria dos comentários submetidos foram positivos, é difícil qualificar como boa a acurácia de 82% nas avaliações de usuários. Para confirmar que existe um problema na classificação de textos negativos seria necessário avaliar a classificação de um volume maior de textos negativos com o modelo. Confirmado o problema, como soluções poderiam ser aplicados hiper-parâmetros ou adicionar mais dados de treinamento para os classificadores de análise de sentimento.

Mesmo com estas observações, o método conseguiu identificar tópicos e classificar sentimentos. Os testes com usuários ajudaram a comprovar que existe coerência na organização dos textos pelos tópicos gerados com LDA. Dentre os textos analisados apenas 9% não tinham qualquer relação com os 3 primeiros tópicos sugeridos. Estes resultados indicam que o modelo proposto pode ser utilizado em sistemas reais.

Como trabalhos futuros, as mesmas técnicas poderiam ser utilizadas para analisar redes sociais, detectando polaridades de comentários de usuários sobre produtos e serviços. Indo mais além, descobrindo o viés político de usuários sobre determinados candidatos durante as eleições - extraindo os principais temas discutidos durante o pleito, ou até mesmo associando assuntos a candidatos. A metodologia aplicada neste trabalho pode servir como base para estas futuras análises sugeridas.

OTTO: UNVAIL SUBJECTS AND SENTIMENTS IN COMMENTS

Abstract: This paper presents a brief study in text mining techniques and explore their applications in Portuguese texts. This paper proposes a model capable of analyzing texts by extracting subjects and detecting the author's sentiment about the extracted subject. The proposed model was apply to texts captured from TripAdvisor, evaluating restaurants. LDA was apply for topic modeling while SVM and Naive Bayes were used to sentiment analysis. For comparative purposes, a simple approach to that uses word dictionary and lexicals was proposed. The machine learning approach was able to identify subjects and sentiments better than our simple one approach, but did not perform well when tested with real users.

Keywords*:* text mining, sentiment analysis, text classification, topic extraction, topic modeling, lda.

REFERÊNCIAS

ALAMSYAH, A. et al. Dynamic large scale data on twitter using sentiment analysis and topicmodeling. **In:2018 6th International Conference on Information and CommunicationTechnology (ICoICT)**. [S.l.: s.n.], 2018. p. 254–258.

AZIZ, M. N. et al. Sentiment analysis and topic modelling for identification of government service satisfaction. **In: 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)**. [S.l.: s.n.], 2018. p. 125–130.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435.

FALEIROS, T. d. P.; LOPES, A. d. A. **Modelos probabilísticos de tópicos: desvendando o latent Dirichlet allocation**. [S.l.]: ICMC-USP, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. Springer, 2009. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.

İŞGÜDER-ŞAHIN, G. G.; ZAFER, H. R.; ADAH, E. Polarity detection of turkish comments on technology companies. In: 2014 International Conference on Asian Language Processing (IALP). [S.l.: s.n.], 2014. p. 136–139

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. [S.l.: s.n.], 2008. v. 2.

KIM, S. N.; CAVEDON, L. Classifying domain-specific terms using a dictionary. **In: Proceedings of the Australasian Language Technology Association Workshop 2011**. Canberra, Australia: [s.n.], 2011. p. 57–65. Disponível em: <https://www.aclweb.org/anthology/U11-1009>.

LU, F. et al. A method of sns topic models extraction based on self-adaptively lda modeling. **In: 2013 Third International Conference on Intelligent System Design and Engineering Applications**. [S.l.: s.n.], 2013. p. 112–115.

NIGAM, K. et al. Learning to classify text from labeled and unlabeled documents. **In: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence**. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998. (AAAI '98/IAAI '98), p. 792–799. ISBN 0-262-51098-7. Disponível em: <http://dl.acm.org/citation.cfm?id=295240.295806>.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? sentiment classification using machine learning techniques. **In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)**. Association for Computational Linguistics, 2002. p. 79–86. Disponível em: <https://www.aclweb.org/anthology/W02-1011>.

PRABOWO, F.; PURWARIANTI, A. Instagram online shop's comment classification using statistical approach. **In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)**. [S.l.: s.n.], 2017. p. 282–287.

RANE, A.; KUMAR, A. Sentiment classification system of twitter data for us airline serviceanalysis. **In:2018 IEEE 42nd Annual Computer Software and Applications Conference(COMPSAC)**. [S.l.: s.n.], 2018. v. 01, p. 769–773. ISSN 0730-3157.

SALINCA, A. Business reviews classification using sentiment analysis. [S.l.: s.n.], 2015. p. 247–250.

SARAKIT, P. et al. Classifying emotion in thai youtube comments. **In: 2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)**. [S.l.: s.n.], 2015. p. 1–5.

SHARMA, R.; NIGAM, S.; JAIN, R. Opinion mining of movie reviews at document level. **International Journal on Information Theory**, v. 3, 08 2014.

TANG, W. et al. A topic label extraction method for the university BBS. In: 2016 **IEEE FirstInternational Conference on Data Science in Cyberspace (DSC).** [S.l.: s.n.], 2016. p.678–682.

ZVAREVASHE, K.; OLUGBARA, O. O. A framework for sentiment analysis with opinion mining of hotel reviews. **In: 2018 Conference on Information Communications Technology and Society (ICTAS)**. [S.l.: s.n.], 2018. p. 1–4.