UNIVERSIDADE FEDERAL DO ACRE
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ANÁLISE DE ALGORITMOS SUPERVISIONADOS NA TAREFA DE CLASSIFICAÇÃO DA POLARIDADE DE REVISÕES

RIO BRANCO
2018

VITOR HUGO DA SILVA LIMA

ANÁLISE DE ALGORITMOS SUPERVISIONADOS NA TAREFA DE CLASSIFICAÇÃO DA POLARIDADE DE REVISÕES

Monografia apresentada como exigência final para obtenção do grau de bacharel em Sistemas de Informação da Universidade Federal do Acre.

Prof. Orientador: Raoni Simões Ferreira, Dr.

RIO BRANCO

2018

Ficha catalográfica elaborada pela Biblioteca Central da UFAC

L732a Lima, Vitor Hugo da Silva, 1997-
Análise de algoritmos supervisionados na tarefa de classificação da popularidade de revisões / Vitor Hugo da Silva Lima. – 2018.
52 f. : il. ; 30 cm.

Monografia (Graduação) – Universidade Federal do Acre, Centro de Ciências Centro de Ciências Exatas Tecnológicas, Curso de Sistemas de Informação. Rio Branco, 2018.
Inclui referências bibliográficas e apêndices.
Orientador: Prof. Dr. Raoni Simão Ferreira.

1. Sistema de informação. 2. Algoritmos - Análise. 3. Algoritmos – Aprendizados de máquinas. 4. Naive Bayes. 5. Support Vector Machine (SVM). 6. Regressão logística. I. Título.

CDD: 004

Bibliotecária: Vivyanne Ribeiro das Mercês Neves CRB-11/600

TERMO DE APROVAÇÃO

VITOR HUGO DA SILVA LIMA

ANÁLISE DE ALGORITMOS SUPERVISIONADOS NA TAREFA DE CLASSIFICAÇÃO DA POLARIDADE DE REVISÕES

Esta monografia foi apresentada como trabalho de conclusão de Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Acre, sendo aprovado pela banca constituída pelo professor orientador e membros abaixo mencionados.

Compuseram a banca:

Prof. Raoni Simões Ferreira, Dr.
Curso de Bacharelado em Sistemas de Informação

Prof. Luiz Augusto Matos da Silva, Dr.
Curso de Bacharelado em Sistemas de Informação

Prof. Olacir Rodrigues Castro Junior, Dr.
Curso de Bacharelado em Sistemas de Informação

Rio Branco, 16 de março de 2018.

À minha família, aos meus amigos e a todos os meus professores.

AGRADECIMENTOS

Agradeço primeiro a Deus por ter me dado a força necessária para concluir mais esta etapa da minha vida.

Agradeço à minha família, em especial aos meus pais e ao meu irmão, pela minha educação, por serem fontes de inspiração e exemplos de comprometimento, integridade e perseverança. Obrigado pelo apoio e incentivo durante meus momentos de maior necessidade, vocês são e sempre serão meus grandes exemplos de vida.

Agradeço a todas as pessoas maravilhosas que tive a oportunidade de conhecer na graduação, em especial, aqueles que se tornaram meus grandes Álvaro Rios e Salatiel Soares.

Um agradecimento ainda mais especial aos meus amigos Alisson Matheus Silva do Vale, Ítalo Rogério de Oliveira Barbosa e Patrick Thanus Mota Batista por terem sido meus companheiros durante esses quatro anos de graduação.

À minha namorada, agradeço por seu apoio e suporte no dia-a-dia foram essenciais em momentos de ansiedade.

Agradeço, em especial, ao professor Dr. Raoni Simões Ferreira, meu orientador, pela dedicação, apoio e paciência durante o desenvolvimento deste trabalho.

"Suba o primeiro degrau com fé. Não é necessário que você veja toda a escada. Apenas dê o primeiro passo."

(Marthin Luther King)

RESUMO

A compreensão de grande volume de conteúdo textual pode significar vantagens para as organizações. Entretanto, analisar conteúdo textual não estruturado de forma manual é uma tarefa dispendiosa e apresenta desafios. Um tipo de análise de interesse desse conteúdo consiste em determinar a polaridade da opinião do autor em relação ao assunto em discussão. Por essa razão, alguns estudos empíricos vêm sendo realizados para determinar quais ferramentas, abordagens ou algoritmos atingem os melhores resultados em diferentes domínios. Este trabalho se propõe a realizar uma análise empírica do comportamento de três algoritmos de aprendizado de máquina supervisionado - *Naive Bayes*, *Support Vector Machine* (SVM) e Regressão Logística - em dois domínios de revisões - revisões de filmes e revisões de hotéis, utilizando as abordagens de atributo-valor tais como presença do termo, frequência do termo e $tf \times idf$. Os resultados foram avaliados usando as métricas de Precisão, Revocação e F1. Os resultados obtidos mostraram um comportamento consistente dos algoritmos SVM e Regressão Logística nas duas coleções utilizadas para as diferentes métricas de desempenho e estratégias de atributo-valor. Em contrapartida, o algoritmo *Naive Bayes* apresentou pior desempenho quando comparado aos demais.

Palavras-chave: Mineração de texto, análise de sentimento, classificação de revisões, aprendizado de máquina, polaridade.

ABSTRACT

The comprehension of large amounts of textual data can provide advantages to organizations which are interested in improving their services and products. However, the task of analyzing unstructured textual contents manually is expensive and presents challenging. One type of interesting analysis of this content is to determine the polarity of an author's opinion on the subject under discussion. For these reasons, some empirical studies have been conducted to determine which tools, approaches or algorithms can achieve the best results in different review domains. This work conducts an empirical analysis of the performance of three supervised machine learning algorithms - Naive Bayes, Support Vector Machine (SVM) and Logistic Regression, in two review domains - movie and hotel reviews, by using three attribute-value-based strategies such as term presence, term frequency and $tf \times idf$. The evaluation was measured considering three performance evaluation metrics - Precision, Recall and F1. The experimental results demonstrate that the Support Vector Machine and Logistic Regression algorithms showed a consistent performance. On the other hand, the Naïve Bayes algorithm showed the worst performance when compared to the others.

Key-words:

Text mining, sentiment analysis, reviews classification, machine learning, polarity.

LISTAS DE FIGURAS

LISTAS DE QUADROS

LISTAS DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Analisar conteúdo textual não estruturado disponível na *web* em *sites*, fóruns, páginas de notícias ou redes sociais, é uma tarefa dispendiosa e apresenta desafios. Compreender de alguma forma esse conteúdo disponibilizado pode significar vantagens para as organizações. Entretanto, é inviável realizar uma análise manual do conteúdo publicado em virtude do grande volume de dados. Um tipo de análise de interesse desse conteúdo consiste em determinar a polaridade da opinião do autor em relação ao assunto em discussão, o que também é conhecido por análise de sentimentos ou polaridade (LIU, 2012). Um exemplo desse tipo de análise é inferir se, em um texto sobre um produto, o autor do texto emite uma opinião favorável, neutra ou desfavorável em relação ao produto.

Opiniões influenciam o comportamento das pessoas. Decisões simples como qual produto comprar, qual filme assistir ou qual hotel se hospedar eram frequentemente tomadas com base nas opiniões de pessoas próximas como por exemplo amigos, especialistas no assunto ou a partir de estudos feitos por institutos especializados. Contudo, a popularidade de conteúdo de opinião, principalmente aqueles veiculados em mídias sociais mudou este cenário, tornando-os mais acessíveis aos usuários comuns e organizações de forma diversificada e em grande volume.

Desta forma, análise de sentimentos tem sido empregada em diversos domínios de aplicações (SADEGH; IBRAHIM; OTHMAN, 2012). Por exemplo, análise

de sentimentos é útil para inferir automaticamente a opinião expressa por um cliente sobre um certo produto de uma loja virtual com base em um comentário de autoria dele. A compreensão da opinião expressa no comentário permite que empresas de comércio eletrônico aprimorem ainda mais seus serviços, produtos e o relacionamento com o cliente. Um domínio tradicional de aplicação da análise da polaridade é o de revisão de filmes (WHITELAW; GARG; ARGAMON, 2005) onde a tarefa é entender as críticas do público sobre o filme a partir dos comentários dos usuários. Outro domínio de interesse é a análise de opiniões de usuários em redes sociais. Todo conteúdo de opinião veiculado em redes sociais pode ser valioso para empresas interessadas em saber a reputação dos seus serviços ou produtos, podendo ter um efeito prático na qualidade dos mesmos. Governos também podem fazer uso das redes sociais para entender a visão do público sobre questões sociais, podendo agir de forma eficaz.

Uma característica comum nos domínios acima citados é que em geral o conteúdo expressa a opinião de um autor sobre uma única entidade e, associada à opinião é dada uma nota. Por exemplo, uma crítica a respeito de uma câmera fotográfica comprada em uma loja virtual pode ser classificada, de acordo com o grau de intensidade da crítica, por meio de uma nota que pode variar de 1 a 5, onde as notas 1 e 2 podem significar uma crítica de sentimento negativo, 4 e 5 positivo e 3 uma opinião neutra. Muito embora, não seja difícil imaginar opiniões acerca de múltiplas entidades, por exemplo, a mesma crítica se referindo a dois modelos de câmera fotográficas, ou a mesma crítica expressando a opinião de dois ou mais autores. Entretanto, neste trabalho, o foco de estudo é voltado para análise de polaridade sobre opiniões expressas por um único autor sobre uma única entidade.

1.1 PROBLEMA DA PESQUISA

Em virtude da aplicabilidade de análise de sentimentos em diversos domínios, há um interesse em particular por soluções automáticas capazes de compreender as opiniões e determinar a sua polaridade de forma rápida e eficiente. Entretanto,

opiniões são escritas em linguagem natural e não são compreendidas por uma máquina se não forem adequadamente tratadas e estruturadas. Por outro lado, a disponibilização de uma ampla literatura relacionada às técnicas e algoritmos para tratamento do texto e aprendizado de padrões de opiniões e, a disponibilização de coleções de opiniões para experimentação e de ferramentas gratuitas que auxiliam a lidar com particulares dessa aplicação, podem fornecer subsídios adequados para realização de estudos sobre polaridade.

Além disso, ao estudar o problema de análise de polaridade, notou-se que é comum abordá-lo como um problema similar ao de classificação de documentos textuais, onde pretende-se atribuir um rótulo ou classe para o documento (LIU, 2010). No contexto de polaridade, o objetivo é a classificação de um documento com base nas palavras que o compõe, isto é, se o autor expressa uma opinião sobre uma entidade com sentimento mais favorável, neutro ou mais negativo. Vários algoritmos baseados em aprendizado de máquina supervisionado têm sido propostos e empregados para a tarefa de classificação automática de polaridade, como em Ortigosa, Martín e Carro (2014); Pang, Lee e Vaithyanathan (2002); Go, Bhayani e Huang (2009); Dave, Lawrence e Pennock (2003). Estes trabalhos são brevemente sumarizados na seção de trabalhos relacionados.

Neste trabalho, procurou-se estudar algoritmos classificação baseado em aprendizado de máquina supervisionado usados para automatizar a tarefa de análise de sentimentos e avaliar o impacto sobre coleções de opiniões. Para tanto, foram usadas duas coleções de opiniões no idioma Inglês empregadas nos trabalhos relacionados de Pang e Lee (2004) e Wang, Lu e Zhai (2010): uma coleção sobre revisões de filmes e uma coleção de comentários sobre hotéis, respectivamente.

O estudo foi orientado pela seguinte questão de pesquisa e que se pretende responder ao final deste trabalho:

- *Um mesmo algoritmo de aprendizado de máquina pode apresentar desempenho consistente quando aplicado a diferentes domínios para diferentes tipos de métricas de qualidade?*

1.2 OBJETIVOS

Este trabalho tem por objetivo principal investigar, testar, avaliar e comparar algoritmos baseados em aprendizado de máquina para identificação automática da polaridade das revisões de múltiplos usuários sobre uma entidade. Em outras palavras, o interesse da pesquisa é estudar algoritmos que analisam se as opiniões dos usuários em comentários expressam sentimentos positivos ou negativos a respeito de uma entidade de interesse *(e.g.*, um produto ou serviço).

Para atingir o objetivo geral da pesquisa, será preciso primeiramente alcançar alguns objetivos específicos, são eles:

a) Modelar o problema de identificação da polaridade como problema de classificação;

b) Explorar um conjunto de abordagens de atributo-valor comumente usados na literatura para caracterizar como os usuários se expressam ao comentar sobre uma entidade usando técnicas de mineração de textos;

c) Determinar a importância desse conjunto de abordagens de atributo-valor sobre o modelo de polaridade;

d) Definir as coleções de revisões de múltiplos autores que serão avaliadas;

e) Definir os algoritmos de aprendizado de máquina supervisionado e avaliar os modelos de polaridade gerados por eles sobre as coleções definidas;

f) Reportar os resultados obtidos dessas avaliações.

1.3 JUSTIFICATIVA

A quantidade de opiniões expressadas pelas pessoas teve um grande aumento com a popularização da internet e do comércio eletrônico. As empresas de comércio eletrônico passaram a encorajar os consumidores a realizarem revisões de seus produtos, isto não ajuda somente aos consumidores no processo decisão da compra de um produto, mas também evita que as empresas precisem gastar dinheiro conduzindo pesquisas ou contratar consultores externos com o objetivo de descobrir a opinião dos seus consumidores a respeito dos seus produtos.

O processo de descoberta de dados é realizado com a aplicação de algoritmos de aprendizado de máquina. O resultado de um algoritmo pode apresentar variações de acordo com o domínio em que é aplicado, por isso é de interesse o estudo de quais algoritmos apresentam resultados mais consistentes quando aplicados a diferentes domínios. Com base nesses resultados o cientista de dados pode justificar a aplicação de um algoritmo em uma base de dados.

1.4 METODOLOGIA

A realização desta pesquisa ocorrerá em 6 etapas, são elas:

a) Revisão bibliográfica em aprendizado de máquina para identificação automática da polaridade em textos de comentários.

b) Seleção de algoritmos de aprendizado de máquina supervisionados que apresentam resultados consistentes em classificação de texto.

c) Obtenção de duas coleções de textos de comentários de domínios diferentes.

d) Definição das métricas de avaliação.

e) Realização de testes nas coleções obtidas com os algoritmos selecionados.

f) Apresentação dos resultados obtidos

1.5 ORGANIZAÇÃO DA MONOGRAFIA

Esta monografia está organizada em quatro capítulos, incluindo este de introdução. O capítulo 2 descreve a fundamentação teórica de análise de sentimentos para o entendimento deste trabalho além de fornecer os trabalhos relacionados a esta monografia. O Capítulo 3 descreve a metodologia seguida para a realização do estudo e em seguida é apresentado o estudo de caso realizado. Ao final, serão apresentadas discussões dos resultados obtidos. O Capítulo 4 descreve as conclusões obtidas e os trabalhos futuros a serem realizados.

2 ANÁLISE DE SENTIMENTO

A análise de sentimentos, também chamada de mineração de opinião, é o estudo computacional das opiniões, sentimentos e emoções expressas em textos a respeito de entidades como produtos, serviços, organizações, indivíduos e seus atributos (GUPTA; LEAL, 2009). Segundo Liu (2012), outros termos podem ser usados tais como como extração de opinião, mineração de sentimento, análise subjetiva, análise afetiva, análise de emoções. Mas, em geral, todos eles se referem aos termos *análise de sentimentos* e *mineração de opinião*.

O cenário mais comum usado para analisar sentimentos é o de revisões de produtos de uma loja virtual, onde revisões são conteúdos gerados por usuários e demonstram a opinião ou sentimento sobre um produto. Tais textos têm chamado atenção por ser uma fonte de valor comercial para empresas interessadas em aprimorar seus produtos e serviços com base na opinião dos usuários (GUPTA; LEAL, 2009), o que acarreta em retorno financeiro, e por ser uma rica fonte de estudos para áreas humanas tais como psicologia, sociologia a fim de entender o comportamento e atitudes pessoais dos usuários online (TANG; TAN; CHENG, 2009). Entretanto, tem sido difícil para essas companhias e especialistas analisarem a enorme massa de revisões disponíveis online para obter, por exemplo, as últimas tendências, para sumarizar ou para inferir opinião geral sobre produtos devido a diversidade desse conteúdo de opinião. Além disso, tais textos são escritos em linguagem natural o que

torna necessária a automação da extração, do tratamento e do entendimento das opiniões expressas nessas revisões.

Técnicas de mineração de texto (MT) têm sido empregados neste tipo de análise para descobrir padrões relevantes em dados não estruturados como é o caso de textos de opinião (GUPTA; LEAL, 2009). Para que essas técnicas de mineração tenham o efeito esperado, as revisões precisam ser primeiramente pré-processadas com a intenção de prepará-las para serem usadas na tarefa de classificação de polaridade. Nas seções a seguir, são descritas como o processamento de linguagem natural pode ser empregado e os procedimentos necessários para realizar da mineração do texto de opinião.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de linguagem natural é o conjunto de técnicas computacionais para análise automática e representação da linguagem humana. A análise automática de texto envolve profundamente o entendimento da linguagem natural pelas máquinas (GUPTA; LEHAL, 2009). Computadores possuem um grande limitador, e isso se deve ao fato de que eles apenas entendem conteúdo que está explícito no texto, ao contrário de nós, processadores de textos humanos, computadores não conseguem assimilar conceitos semânticos relacionados, desambiguação, envolvimento textual, etiquetagem de função semântica[1] (CAMBIRA; WHITE, 2014).

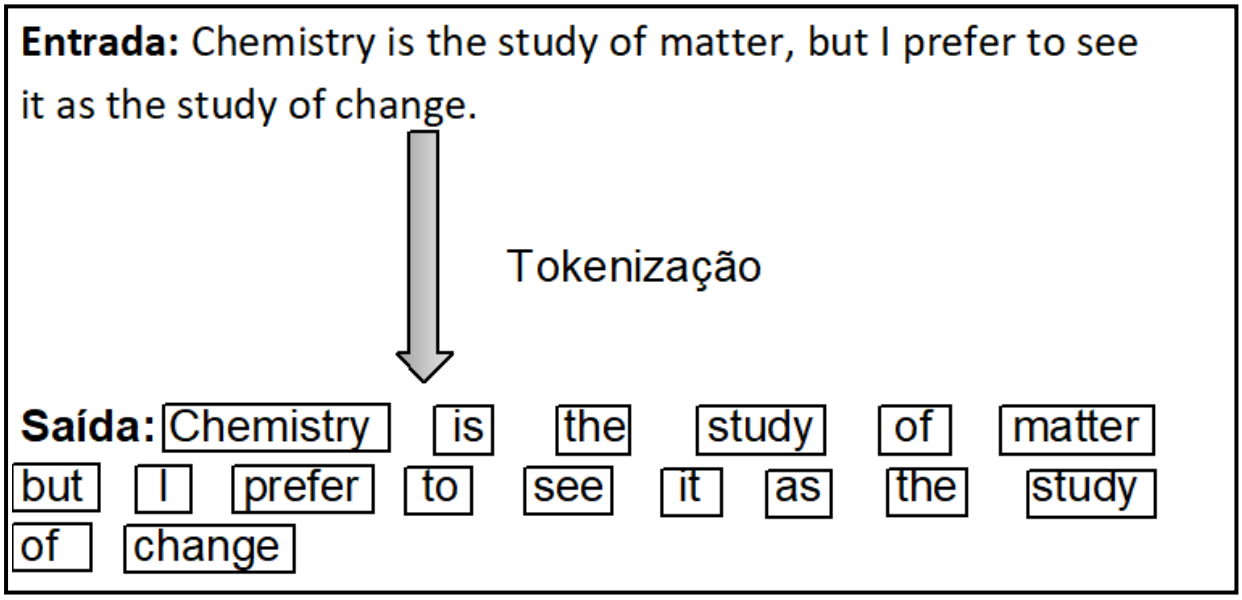
Para o pré-processamento das revisões são utilizadas algumas técnicas de processamento de linguagem natural, dentre elas a tokenização e a remoção de *stop words*[2]. Estas técnicas são detalhadas nas subseções a seguir.

[1] A etiquetagem de função semântica refere-se à categorização de uma palavra de acordo com sua orientação semântica (*e.g.*, estudar-verbo).
[2] Palavras que não carregam nenhuma informação (*e.g.*, preposições).

2.1.1 Tokenização

Dado uma sequência de caracteres s e um determinado documento de revisão d, a tokenização refere-se a tarefa de separar e armazenar palavras em estruturas chamadas *tokens* e ao mesmo tempo eliminar certos caracteres, como pontuação (CAMBRIA; WHITE, 2014). A Figura 1 representa um exemplo de tokenização de um documento, onde um texto é fornecido como entrada e o processo de tokenização gera uma saída do mesmo documento com as palavras armazenadas em *tokens* e sem pontuação.

Figura 1: Tokenização de um documento.



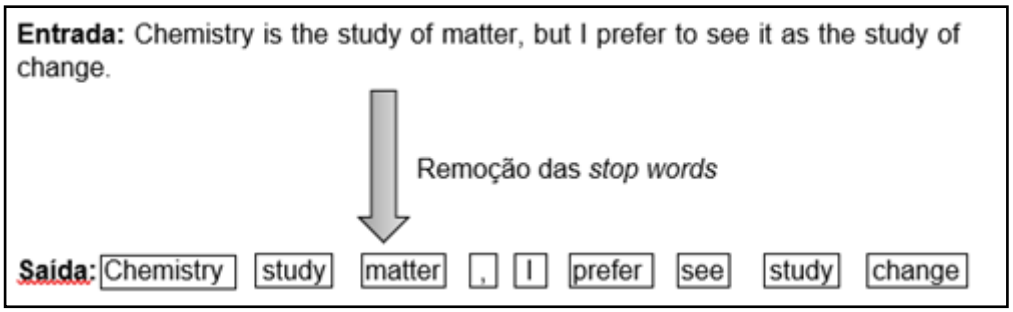
Fonte: Elaboração própria.

2.1.2 STOP WORDS

Stop words são palavras altamente frequentes que não carregam nenhuma informação, como por exemplo, pronomes, preposições, conjunções, etc. A remoção das *stop words* pode aprimorar os resultados em classificação de texto e reduzir a complexidade computacional (GUPTA; LEHAL, 2009). É importante ressaltar que o

processo remoção das *stop words* ocorre logo depois da tokenização. No exemplo da Figura 2 é possível notar que as palavras *is, the, but, to, it, as* e *of* são removidas da sentença.

Figura 2: Remoção das stop words.



Fonte: Elaboração própria.

2.2 O PROCESSO DE MINERAÇÃO DE TEXTOS

O processo de mineração de texto é semelhante ao processo de mineração de dados (MD). Porém, enquanto MD trabalha com dados estruturados, o processo de MT trabalha com dados não estruturados, geralmente na forma de texto ou documentos, havendo, portanto, um tratamento diferenciado em algumas etapas (MARTINS et al., 2003).

A tarefa de classificação de polaridade está relacionada ao processo de MT, pois é uma forma de extração de conhecimento utilizando classificação. Segundo Martins et al. (2003) o processo de MT está dividido em quatros etapas: (a) coleta de documentos, (b) pré-processamento, (c) extração do conhecimento e (d) avaliação e interpretação dos resultados. A Figura 3 mostra o fluxo dessas etapas.

Figura 3: Etapas do processo de mineração de texto.

Fonte: Elaboração própria.

As etapas de MT são descritas resumidamente da seguinte forma:

- A coleta de documentos é o processo onde os documentos relacionados com o domínio da aplicação são coletados a partir de um repositório;
- Na etapa de pré-processamento acontecem uma série de ações, como aplicações de técnicas de processamento de linguagem natural e a transformação dos documentos em um formato atributo-valor, são realizadas sobre o conjunto de documentos com o objetivo de tornar possível a aplicação dos algoritmos de aprendizado e a realização do processo de extração de conhecimento (AAS; EIKVIL, 1999);
- Na etapa de extração de conhecimento ocorre a aplicação de algoritmos de aprendizado de máquina com a finalidade de extrair conhecimento em formas de classificação, regras de associação, relações, regressão, entre outros;
- Na etapa de avaliação e interpretação dos resultados, ocorre a avaliação do desempenho do algoritmo classificador, esta etapa pode ser realizada com o auxílio de métricas como acurácia, revocação, precisão.

2.2.1 Coleta e pré-processamento

A coleta de documentos é realizada a partir de um ou vários repositórios. Com uma grande quantidade de documentos coletados, há um desafio significativo para organizar, remover conteúdos irrelevantes e entender o conteúdo desses documentos.

Os documentos coletados podem estar em diferentes padrões e contendo conteúdo irrelevante. Os documentos são padronizados e o conteúdo irrelevante é removido, então esses documentos são armazenados em uma base de documentos. Após a coleta, é necessário um pré-processamento desses documentos preparando-os para serem representados em uma forma adequada (c.f Seção 2.1) para, posteriormente, aplicar técnicas de extração do conhecimento, utilizando sistemas de aprendizado de máquina, com a finalidade de descobrir padrões úteis ou desconhecidos presentes nos documentos. A representação de documentos mais comum é descrevê-los por meio de um conjunto de palavras ou termos que ocorrerem no documento.

2.2.1.1 Representação dos documentos usando a tabela atributo-valor

O procedimento adotado para representação de cada documento de revisão d_i é descrevendo-o como um vetor de m termos que ocorrem no documento. Um termo pode ser representado por uma palavra simples (palavra de tamanho 1, também chamadas de 1-grama ou unigrama) ou por uma composição de palavras (n-gramas). Cada termo, portanto, será um elemento do conjunto de atributos da tabela atributo-valor. Espera-se que ao final desta transformação, a coleção de documentos possa ser representada conforme ilustrado no Quadro 1 a seguir:

Quadro 1. Representação atributo-valor.

	t_1	t_2	...	t_j
d_1	a_{11}	a_{12}	...	a_{1j}
d_2	a_{21}	a_{22}	...	a_{2j}
...	...	...	...	...
d_i	a_{i1}	a_{i2}	...	a_{ij}

Fonte: (Martins *et al*, 2003)

O Quadro 1 representa d documentos (instâncias) e i termos (atributos), e cada documento d_i é uma tupla $d_i = (a_{i1}, a_{i2}, ..., a_{ij})$. O valor a_{ij} refere-se ao valor associado ao j-ésimo termo do documento i, ou seja, a_{ij} representa o valor do termo t_j no documento d_i e pode ser calculado de diversas formas (Martins et al, 2003).

Para calcular o valor de a_{ij} pode-se, por exemplo, usar valores binários para significar a presença do termo (do inglês, *term presence*) j no documento i (tp_{ij}), neste caso a_{ij}=1, e a_{ij}=0 para significar a ausência do termo j. Outras medidas estatísticas também podem ser consideradas, como por exemplo, a frequência do termo (do inglês, *term frequency*) (tf_{ij}), que considera o número de ocorrências do termo j em um documento d_i . A frequência do termo é calculada como a frequência absoluta do termo no documento. O valor dos termos pode ser calculado também levando em consideração, além da frequência de um termo, o fator relacionado a frequência inversa do documento (do inglês, *index document frequency)* (idf), favorecendo por sua vez termos que aparecem em poucos documentos de uma coleção, chamada de medida $tf \times idf$ (MANNING; RAGHAVAN; SHUTZE, 2009).

O Quadro 2 descreve as três maneiras usadas para calcular o valor de a_{ij}. Neste quadro as notações N designa o número de documentos da coleção, n_j o número de documentos da coleção que aparece o termo j e f_{jd} a frequência absoluta do termo j em um documento d.

Quadro 2. Abordagens para representação no formato atributo-valor.

Nome	Formula matemática
tf_{ij}	$a_{ij} = f_{jd}$
tp_{ij}	$a_{ij} = \begin{cases} 1, \ se \ tf_{ij} > 0 \\ 0, \ se \ tf_{ij} = 0 \end{cases}$
$tf \times idf$	$a_{ij} = \ f_{jd} \times \log\left(\frac{N+1}{n_j+1}\right)$

Fonte: Elaboração própria

Um exemplo de atribuição do valor a_{ij}utilizando as medidas anteriores é ilustrado a seguir, utilizando a base de dados de revisões fictícia do Quadro 3.

Quadro 3. Base de dados fictícia.

ID	Conteúdo
#1	The movie is really really good
#2	Horrible movie
#3	Waste of time

Fonte: Elaboração própria

Considerando a presença do termo como valor de atributo, atribui-se valor 1, em caso da frequência do termo seja maior que 0. Caso contrário, atribui-se 0 para o atributo caso o termo não apareça nenhuma vez no documento. Um exemplo de como seria representado o documento usando a ausência ou presença do termo j no documento d_i é mostrado no Quadro 4.

Quadro 4. Atributo-valor definido pela presença do termo na revisão.

#	horrible	the	waste	good	is	movie	of	really	Time
1	0	1	0	1	1	1	0	1	0
2	1	0	0	0	0	1	0	0	0
3	0	0	1	0	0	0	1	0	1

Fonte: Própria

A frequência do termo é medida pela quantidade de vezes que o termo j aparece em um documento d_i. Neste caso, $a_{ij} = f_{jd}$. Um exemplo pode ser visto no Quadro 5.

Quadro 5. Atributo-valor definido pela frequência absoluta do termo na revisão.

#	horrible	the	waste	good	is	movie	of	really	Time
1	0	1	0	1	1	1	0	2	0
2	1	0	0	0	0	1	0	0	0
3	0	0	1	0	0	0	1	0	1

Fonte: Própria

A abordagem $tf \times idf$, é uma estratégia que busca assinalar um valor, a_{ij} para o termo j em um documento d_i como o produto de um fator do número de ocorrências do termo j no documento pela frequência inversa do número de ocorrências do termo j na coleção N, onde o atributo ocorre no mínimo uma vez (MANNING; RAGHAVAN; SHUTZE, 2009). De acordo com o exemplo ilustrado da base de dados do Quadro 3, usando a abordagem $tf \times idf$, cada documento será representado pelos valores de atributos mostrados no Quadro 6.

Quadro 6. Atributo-valor definido pela medida $tf \times idf$ do termo na revisão.

#	horrible	the	waste	good	is	movie	of	really	Time
1	0	0,36	0	0,36	0,36	0,27	0	0,72	0
2	0,79	0	0	0	0	0,60	0	0	0
3	0	0	0,57	0	0	0	0,57	0	0,57

Fonte: Própria

É importante ressaltar que nos exemplos demonstrados não foi considerado a aplicação de remoção das *stop words*. Após o processo de transformação dos documentos para o formato atributo-valor, é dada continuidade ao processo chamado de extração do conhecimento.

2.2.2 Extração do conhecimento

Nesta etapa é definida a tarefa de aprendizado de máquina a ser utilizada. Neste trabalho a tarefa de aprendizado definida é a de classificação da polaridade de revisões.

Uma vez definida a tarefa, o próximo passo é escolher o algoritmo de aprendizado de máquina. Em classificação de texto, em geral, usa-se algoritmos de aprendizado supervisionados e não supervisionados. Neste trabalho, a tarefa de classificação de polaridade foi modelada como um problema de classificação de texto supervisionado. Portanto, foram selecionados algoritmos de aprendizado de máquina supervisionado para tarefa de classificação. Maiores detalhes serão tratados na subseção 2.3.

2.2.3 Avaliação e interpretação dos resultados

Terminada a etapa de extração do conhecimento, os resultados obtidos devem ser interpretados e avaliados. Nessa última fase os resultados obtidos são analisados por meio de métricas de qualidade de classificadores tais como Precisão, Revocação e F1.

É nesta etapa, onde são avaliados os resultados obtidos e se atendem ao interesse originalmente proposto, que no caso deste trabalho é de avaliar a capacidade de um classificador automático em aprender a distinguir revisões positivas de revisões negativas. De acordo com Han, Kamber e Pei (2012), para entender a realização da etapa de avaliação dos resultados, é preciso estar familiarizados com alguns conceitos, são eles:

a) Verdadeiros positivos (VP): são as instâncias positivas que foram corretamente rotuladas como positivas pelo classificador;

b) Verdadeiros negativos (VN): são as instâncias negativas que foram corretamente rotuladas como negativas pelo classificador;

c) Falsos positivos (FP): são as instâncias negativas que foram incorretamente rotuladas como positivas pelo classificador;

d) Falsos negativos (FN): são as instâncias positivas que foram incorretamente rotuladas como negativas pelo classificador.

Esses termos são resumidos em uma matriz de confusão conforme ilustrada no Quadro 7.

Quadro 7. Matriz de confusão.

Predita/Real	**Positivo**	**Negativo**
Positivo	Verdadeiros Positivos (VP)	Falsos Positivos (FP)
Negativo	Falsos Negativos (FN)	Verdadeiros Negativos (VN)

Fonte: Própria

A matriz de confusão é uma ferramenta para analisar o desempenho do classificador de acordo com diferentes classes. Dado c classes (onde $c > 2$, a matriz de confusão é uma tabela de tamanho mínimo $c \times c$.

É com base na matriz de confusão que é possível calcular os valores das métricas de Precisão, Revocação e F1 (BAEZA-YATES; RIBEIRO-NETO, 2011).

A métrica de Revocação avalia a taxa de instância positivas classificadas corretamente e pode ser entendida de uma maneira mais simples como o total de instâncias positivas preditas corretamente dividido pelo número total de instâncias positivas. Esta medida é expressa pela equação a seguir:

$$\frac{VP}{VP + FN} = \frac{\sum_i^n 1[y_i = 1]}{\sum_i^n 1[y_i = 1]} \tag{2.1}$$

A métrica de Precisão é calculada como demonstrada pela equação (2.2), e pode ser resumida como sendo as instâncias classificadas como positivas corretamente dividido pelo número total de instâncias classificadas pelo classificador.

$$\frac{VP}{VP + FP} = \frac{\sum_i^n 1[y_i = 1]}{\sum_i^n 1[y_i = 1]} \quad (2.2)$$

A métrica F1, expressa na equação (2.3), é dada pela média harmônica entre as métricas de Precisão e Revocação, pode ser lida como a precisão vezes revocação dividido pela Precisão mais Revocação, tudo isso multiplicado por dois. A medida F1 é uma métrica que sumariza as métricas Precisão e Revocação, e é muito usada em classificação como indicador de qualidade do classificador.

$$F1 = 2\frac{precisao \times revocacao}{precisao + revocacao} \quad (2.3)$$

2.3 APRENDIZADO DE MÁQUINA SUPERVISIONADO

Aprendizado de máquina é uma subárea da Inteligência Artificial que se preocupa com o projeto e o desenvolvimento de algoritmos que ensinam as máquinas a utilizar padrões presentes nos dados fornecidos como entrada. O processo de aprendizado de máquina é similar com a forma que nós (humanos) aprendemos por meio da experiência, adaptando o conhecimento previamente adquirido e adaptado para uma nova situação (BAEZA-YATES; NETO, 2011).

Para tornar tal processo de aprendizado factível, alguns algoritmos são projetados com a intenção de permitir que, após uma fase de treinamento sobre um conjunto de instâncias (exemplos) rotuladas por um especialista humano, uma máquina (ou computador) seja capaz de interpretar novas instâncias e classificá-las apropriadamente a partir de uma generalização do que foi apresentado anteriormente. A esse processo de fornecimento de exemplos de entradas para esses algoritmos para uma máquina com objetivo de aprender uma regra geral (também chamada de modelo do aprendizado) é dado o nome de aprendizado supervisionado. Análise de sentimentos pode ser vista como um problema de classificação supervisionado (LIU, 2010) cujo o interesse principal é aprender, a partir de um conjunto de revisões previamente rotuladas, uma regra geral para distinguir a classe das futuras revisões.

Formalmente o problema de classificação pode ser definido como: dado um conjunto de treino (x_i, y_i) para $i = 1, \ldots, n$, onde x representa uma instância (ou seja, um exemplo de revisão) caracterizada por um conjunto de atributos $a_1, .. a_m$ e y_i representa o rótulo da instância x_i. Pretende-se criar uma função de classificação F que seja capaz de predizer um rótulo y para uma nova instância de x, sendo que x pode ser um definido como um vetor de atributos que armazena os valores de uma instância i. Um conjunto de instâncias rotuladas de treinamento, chamada de conjunto de treino, serve de base para desenvolver a função de classificação na qual pode ser usada para fazer predições sobre instâncias novas não rotuladas, conhecidas por conjunto de teste (FERREIRA, 2016).

O Quadro 8 exemplifica a representação de um conjunto de treino, previamente rotulado e usado para construção da função de classificação, e um conjunto de teste, onde serão feitas as previsões das classes e a avaliação do desempenho do classificador.

Quadro 8. Exemplo de base de dados dividida em treino e teste.

		Instância	**Rótulo**
BASE DE DADOS	**Treino**	Todo filme foi muito bom, mas o final foi simplesmente espetacular.	Pos
		Um filme realmente adorável, uma trilha sonora espetacular.	Pos
		Um dos piores filmes de toda história, simplesmente ridículo.	Neg
		Inaceitável que um filme ridículo como esse ganhou um Oscar.	Neg
	Teste	A atriz esteve espetacular durante todo o filme, e a reviravolta que ocorre no meio do filme é espetacular	

Fonte: Própria

Note que, embora o conjunto de teste não deva vir rotulado em um cenário real, é importante conhecer a classe das instâncias de testes quando o propósito é avaliar a capacidade de generalização do classificador.

2.3.1 Algoritmos

O algoritmo *Naive Bayes* (NB) é um algoritmo de classificação baseado no teorema de Bayes e pode ser usado tanto para modelos diagnósticos[3] quanto para preditivos[4]. O nome deriva do fato de utilizar técnicas bayesianas e não levar em conta as dependências entre os atributos que possam existir. Ele projeta um classificador com base nas probabilidades incondicionais dos atributos do conjunto de treinamento. De acordo com Zhang (2004) a probabilidade de um exemplo $E = x_1, x_2, \dots, x_n$ pertencer a uma classe c é dada por:

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)} \tag{2.4}$$

Outro algoritmo utilizado é a Regressão Logística (RL), uma forma não linear da regressão linear, para determinar a contribuição de vários fatores para um resultado. Enquanto a regressão linear traça uma função limitada por $(-\infty, +\infty)$, a regressão logística traça uma função limitada por $(0, 1)$ (LEE, 2010). A definição é dada por:

$$P\theta(y|x) = \frac{e^{\theta\prime x}}{1 + e^{\theta\prime x}} \tag{2.5}$$

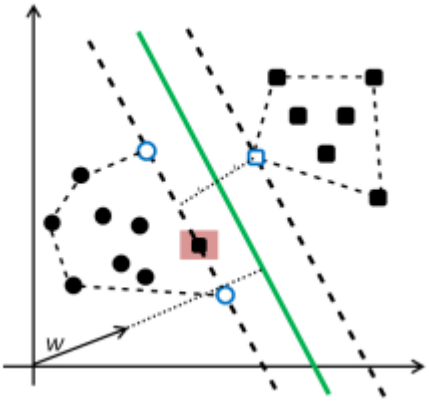
onde $\theta \in R^k$ é um vetor de atributos de dimensão k.

Outro algoritmo de aprendizado de máquina é o *Support Vector Machine* (SVM), que possui por ideia básica traçar hiperplanos de separação com margem máxima entre instâncias das classes existentes. Este hiperplano é aprendido a partir

[3] Objetivo de avaliar impactos de uma ação (*e.g.*, alcance de uma ação de marketing).
[4] Prevê um tipo de comportamento. O que é mais provável que aconteça dado alguns fatores.

do conjunto de treino de um conjunto de treino rotulado. Dentro do hiperplano o SVM modela cada classe como uma região em um espaço vetorial. Os vetores marginais de cada região são utilizados para determinar a margem de separação entre as classes (SILVA, 2015). A Figura 4 ilustra duas classes de documentos e sua representação no espaço. A linha verde indica o hiperplano de separação gerado pelo vetor w. Os vetores de suporte estão representados pelos círculos e retângulos azuis existentes nas linhas pontilhadas. Note que um documento da classe $C_b$está localizado do lado esquerdo do hiperplano, o lado direito do hiperplano corresponde a C_a.

Figura 4: Exemplos de duas classes de documentos no espaço.



Fonte: Silva (2015, p.12).

Então, para a classificação de novos documentos, estes são projetados no hiperplano e observa-se sua relação ao hiperplano de separação. É importante ressaltar que quanto mais próximo o documento é projetado ao hiperplano gerado por w mais difícil é classificá-lo (SILVA, 2015).

2.4 TRABALHOS RELACIONADOS

Nesta seção, é feito um breve resumo dos principais trabalhos relacionados a esta pesquisa e que possuem grande importância na área de análises de sentimentos.

Bernardini, De Castro Lunardi e Viterbo (2015) apresentam um trabalho com foco no levantamento do uso das técnicas de aprendizado supervisionado que são mais úteis na definição de modelos para a classificação de opinião e das aplicações que fazem uso desses algoritmos em diferentes áreas do conhecimento. Zhou e Chaovalit (2005) examinam a mineração de texto utilizando as abordagens de aprendizado de máquina e orientação semântica. Liu e Hu (2004) utilizam a mineração de texto para sumarização de revisões de produtos, detecção das características que os consumidores expressaram opinião, identificação sentenças de opinião em cada revisão e classificação da opinião expressa pelos consumidores. Gupta e Lehal (2009) apresentam um *survey* abordando o processo de mineração de textos e suas áreas de aplicações. Pang, Lee e Vaithyanathan (2002) discorrem sobre o problema de classificação de documentos baseado no sentimento geral demonstrado, por exemplo, determinar se uma revisão é positiva ou negativa utilizando os algoritmos de aprendizado de máquina supervisionado *Naive Bayes, Maximum Entropy e Support Vector Machine*. Go, Bhayani e Huang (2009) também utilizam os algoritmos *Naive Bayes, Maximum Entropy e Support Vector Machine* para classificação automática de textos de mensagens do Twitter, demonstram que há um melhor desempenho quando utilizado dados de emoticons. Ortigosa, Martín e Carro (2013) apresentam um novo método para análise de sentimento na rede social Facebook para, a partir das mensagens publicadas pelos usuários, extrair informações sobre a polaridade do sentimento do usuário transmitida na mensagem, modelar a polaridade do sentimento do mesmo e detectar mudanças significativas no seu humor.

3 ESTUDO DE CASO

Neste capítulo é descrito o estudo de caso de classificação de polaridade em revisões realizado. Em particular, são relatadas as bibliotecas para apoiar a tarefa de mineração de texto e para construção dos classificadores, as características das coleções de revisões, os procedimentos metodológicos seguidos e os resultados obtidos.

3.1 BIBLIOTECAS

As tarefas da etapa de pré-processamento foram feitas utilizando a biblioteca escrita na linguagem Python de código aberto chamada *Natural Language Toolkit* (NLTK)[5]. NLTK foi originalmente criado em 2001 como parte do Departamento de Linguística Computacional no Departamento de Ciência da Computação e Informação da Pensilvânia e é continuamente aprimorada com a ajuda de dezenas de contribuições (BIRD; KLEIN; LOPER, 2009).

NLTK oferece um conjunto de métodos úteis para executar as tarefas de conversão das palavras em minúsculo, de tokenização, de remoção de *stop words* e de extração de atributos textuais tais como frequência do termo e frequência inversa

[5] https://www.nltk.org/

do documento, o que permitiu construir, a partir deles, as três estratégias de atributo-valor usadas neste trabalho - presença do termo, frequência do termo e medida $tf \times idf$. Essas medidas foram extraídas considerando palavras unigramas.

Os experimentos foram feitos usando os algoritmos de classificação disponíveis na biblioteca de código aberto de Aprendizado de Máquina escrita em Python chamada scikit-learn[6]. Foram selecionadas três implementações de algoritmos supervisionados presentes na biblioteca para a tarefa de classificação de polaridade: Support Vector Machine (SVM), Multinomial Naive Bayes (NB) e Regressão Logística (RL).

3.2 BASE DE DADOS

As experimentações foram feitas utilizando duas coleções de revisões no idioma Inglês de domínios diferentes. Uma no domínio de revisões de filmes e a outra no de revisões de hotéis[7]. A primeira coleção é composta por um conjunto de 2000 revisões de filmes com suas respectivas polaridades (cada revisão é rotulada com polaridade positiva ou negativa). Neste conjunto, 50% das revisões são rotuladas com a polaridade positiva e os outros 50% com a polaridade negativa. Esta base de dados foi utilizada no trabalho de Pang e Lee (2004).

A segunda coleção é composta por 9000 revisões de hotéis do site TripAdvisor e foram extraídas da base de dados utilizada por Wang, Lu e Zhai (2010). Diferente das revisões de filmes, as revisões de hotéis apresentam uma nota atribuída pelo autor da revisão. Para rotulação da polaridade das revisões em positivas/negativas foi feito o seguinte procedimento: considerando a nota atribuída pelo autor da revisão, revisões com notas entre 4 e 5 foram rotuladas como revisões positivas, enquanto que revisões com notas entre 0 e 2 foram rotuladas como negativas. As revisões de nota 3 foram consideradas neutras e descartadas da coleção. Como resultado deste procedimento, obteve-se um total de 8000 revisões positivas e 1000 negativas. Optou-

[6] http://scikit-learn.org/stable/
[7] Ambas estão disponíveis em: https://github.com/vdhug/AnaliseDeSentimento

se também por criar uma segunda versão desta coleção onde a quantidade de revisões positivas e negativas fossem as mesmas a partir de uma seleção de amostras aleatórias da coleção original.

Originalmente a coleção de revisões de hotéis possui uma quantidade total de 21.617 palavras diferentes, enquanto a coleção de revisões de filmes possui uma quantidade total de 39.659 palavras diferentes. Após a etapa de pré-processamento, onde termos dessas coleções foram convertidos para minúsculos para posteriormente serem tokenizados e removidas as *stop words*, foi montado o vetor de atributos (palavras) que representam as instâncias (revisões). A representação dessas instâncias foi feita usando as três estratégias de formato atributo-valor: *term presence* (tp) - presença ou ausência da palavra na revisão; *term frequency* (tf) - frequência da palavra e o $tf \times idf$ calculado como o produto frequência da palavra pelo fator da frequência inversa da palavra (c.f. Seção 2.2.1.1).

O Quadro 9 apresenta algumas estatísticas de cada coleção após a etapa de pré-processamento.

Quadro 9. Visão geral das características das coleções.

#	**Revisões de hotéis balanceada (TripAdvisor)**	**Revisões de hotéis desbalanceada (TripAdvisor)**	**Revisões de filmes**
Número de revisões negativas	1000	1000	1000
Número de revisões positivas	1000	8000	1000
Quantidade máxima de palavras por revisão	744	1157	907
Mínimo de palavras por revisão	5	2	6
Média de palavras por revisão	159	112	235

Fonte: Própria

3.3 METODOLOGIA DE EXPERIMENTAÇÃO

A metodologia usada para avaliar o desempenho de algoritmos de aprendizado de máquina foi a validação cruzada com 10 partições. Essa metodologia obtém métricas de desempenho mais confiáveis e é adotada para avaliar a capacidade de generalização do classificador em conjuntos de dados independentes (WITTEN; FRANK, 2016). Neste procedimento, a coleção de revisões foi particionada em 10 subconjuntos de revisões. A partir dos 10 subconjuntos, um único subconjunto é usado como conjunto de teste enquanto que os 9 restantes são usados como conjunto de treino. Esse processo é repetido 10 vezes de tal forma que cada um dos 10 subconjuntos é usado uma única vez como conjunto de teste.

As partições usadas para treinamento e testes foram as mesmas para todos os classificadores. Para cada conjunto de teste avaliado foi obtido o valor das métricas Precisão, Revocação e F1. Por fim, obteve-se a média de toda validação cruzada como a média dos valores obtidos para as 10 rodadas de treino e teste. Também foi garantido que todos os experimentos de treino e teste continham a mesma proporcionalidade de revisões positivas e negativas observada na coleção maior.

Por se tratar de um problema de classificação multi-classes[8] onde há o interesse em classificar corretamente revisões que expressam sentimentos positivos (classe positiva) como as que expressam sentimentos negativos (classe negativa), procurou-se, portanto, reportar o desempenho dos classificadores para as duas classes.

[8] Classificação multi-classes ocorre quando existem duas ou mais classes de interesse, diferente da binária que existe apenas uma classe de interesse (*e.g.*, detecção de e-mails que são spams).

3.4 RESULTADOS

Nesta seção, são reportados os resultados obtidos para os três algoritmos de classificação escolhidos *Naive Bayes* (NB), *Support Vector Machine* (SVM) e Regressão Logística (RL) nas duas coleções de revisões.

Cada algoritmo foi treinado usando as três estratégias formato atributo-valor e identificado no seguinte formato: ALG-TF, ALG-TP e ALG-TF IDF. Por exemplo, NB-TF, NB-TP e NB-TF IDF significa que o algoritmo Naive Bayes foi treinado com instâncias cujos os atributos correspondem a frequência do termo, a presença do termo e a medida de importância $tf \times idf$ do termo, respectivamente. Esse modelo de identificação foi estendido para os demais algoritmos.

A Tabela 1 detalha os resultados obtidos na coleção de revisão de hotéis desbalanceada.

Tabela 1. Resultados gerais aplicados ao domínio de hotéis - coleção desbalanceada.

	F1		Precisão		Revocação	
	Pos	**neg**	**pos**	**neg**	**pos**	**Neg**
NB - TF	0,96	0,67	0,95	0,77	0,97	0,61
SVM – TF	0,96	0,70	0,96	0,73	0,96	0,68
RL – TF	0,96	0,73	0,96	0,80	0,97	0,67
NB – TP	0,96	0,63	0,94	0,82	0,98	0,52
SVM – TP	0,96	0,71	0,96	0,75	0,97	0,68
RL – TP	0,97	0,73	0,96	0,81	0,98	0,67
NB – TF-IDF	0,94	0,0	0,88	0,0	1,0	0,0
SVM – TF-IDF	0,97	0,73	0,95	0,85	0,98	0,65
RL – TF-IDF	0,96	0,55	0,92	0,94	0,99	0,39

Fonte: Própria

De acordo com a Tabela 1, foi possível notar, para todas as métricas, que todos os classificadores apresentaram desempenho superior em classificar revisões positivas do que em classificação de revisões negativas. Os melhores resultados, de acordo com a métrica F1, entre todos os métodos foram alcançados pelos algoritmos SVM-TF IDF e RL-TP (0,97 para revisões positivas e 0,73 para negativas). Já os algoritmos NB-TP, NB-TF e NB-TF IDF apresentaram os piores resultados.

Considerando a medida de precisão dos classificadores observou-se um desempenho estável e melhor para revisões positivas. Com exceção ao apresentado pelo algoritmo NB-TF IDF (0.88), a precisão nas revisões positivas ficou sempre entre 0,94 e 0,96, enquanto nas revisões negativas, a variação foi entre 0,73 e 0,94. Esse mesmo comportamento é percebido na medida de revocação. Onde, com exceção do algoritmo NB-TF IDF, a revocação nas revisões positivas foi entre 0,96 e 0,99 e nas negativas entre 0,39 e 0,68.

Os resultados acima mostram que o desempenho ruim do NB pode ter sido afetado pelo desbalanceamento dos dados da coleção. Por essa razão o classificador NB-TF IDF não conseguiu construir uma função de classificação que detectasse revisões negativas, classificando todas as instâncias como positivas em todos os *folds* de testes.

Para saber o comportamento do NB em um ambiente em que o algoritmo obtivesse a mesma quantidade de instâncias da classe positiva e negativa para aprendizado, foi realizado um balanceamento dos dados da coleção de revisões de hotéis.

A Tabela 2 detalha os resultados obtidos na coleção de revisão de hotéis balanceada.

Tabela 2. Resultados gerais aplicados ao domínio de hotéis - coleção balanceada.

	F1		Precisão		Revocação	
	pos	**neg**	**pos**	**neg**	**pos**	**neg**
NB - TF	0,90	0,87	0,86	0,95	0,96	0,81
SVM – TF	0,90	0,89	0,89	0,90	0,90	0,89
RL – TF	0,91	0,90	0,90	0,91	0,91	0,90
NB – TP	0,90	0,88	0,86	0,95	0,96	0,82
SVM – TP	0,90	0,90	0,90	0,91	0,91	0,89
RL – TP	0,91	0,90	0,90	0,92	0,93	0,89
NB – TF-IDF	0,88	0,82	0,81	0,97	0,96	0,74
SVM – TF-IDF	0,92	0,92	0,91	0,93	0,93	0,90
RL – TF-IDF	0,91	0,91	0,90	0,92	0,93	0,89

Fonte: Própria

De acordo com a Tabela 2, foi possível notar que a precisão na classificação de revisões negativas foi sempre superior em relação as positivas, já com base nas métricas F1 e revocação, todos os classificadores apresentaram desempenho igual ou superior quando trata-se da classificação de revisões positivas, o que não significou que o desempenho em revisões negativas tenha sido ruim. Além disso, levando em consideração a métrica F1, foi notado que o melhor resultado entre todos os métodos foi obtido pelo algoritmo SVM-TF IDF (0,92 para positivas e negativas), seguido pelo algoritmo RL-TF IDF (F1 de 0,91 para positivas e negativas). Já o algoritmo NB-TF IDF apresentou o pior desempenho entre todos (0,88 para positivas e 0,82 para negativas).

O algoritmo SVM-TF IDF também obteve os melhores resultados nas métricas de Precisão (0,91 nas revisões positivas e 0,93 nas negativas) e revocação (0,93 nas revisões positivas e 0,90 nas negativas). Enquanto o algoritmo NB-TF IDF obteve os piores resultados nas métricas de precisão (0,81 nas revisões positivas e 0,97 nas negativas) e revocação (0,96 nas revisões positivas e 0,74 nas negativas).

A Tabela 3 detalha os resultados obtidos na coleção de revisão de filmes.

Tabela 3. Resultados gerais aplicados ao domínio de filmes.

	F1		Precisão		Revocação	
	pos	**neg**	**pos**	**neg**	**pos**	**neg**
NB - TF	0,80	0,81	0,82	0,79	0,78	0,83
SVM – TF	0,83	0,83	0,85	0,82	0,81	0,85
RL – TF	0,84	0,85	0,85	0,84	0,84	0,85
NB – TP	0,82	0,83	0,84	0,81	0,80	0,84
SVM – TP	0,84	0,84	0,85	0,84	0,84	0,85
RL – TP	0,86	0,86	0,86	0,85	0,85	0,86
NB – TF-IDF	0,80	0,82	0,83	0,79	0,78	0,84
SVM – TF-IDF	0,84	0,84	0,83	0,85	0,85	0,83
RL – TF-IDF	0,82	0,82	0,81	0,83	0,83	0,81

Fonte: Própria

No domínio de revisões de filmes ocorre uma inversão da performance dos algoritmos classificadores, com base na medida F1. Na maioria dos casos houve um empate na classificação de revisões positivas e negativas. Em outros, os classificadores apresentaram um desempenho superior para revisões negativas em relação às positivas.

A avaliação com base nas métricas de desempenho demonstrou que a medida F1 máxima obtida por um método foi através do algoritmo RL-TP (0,86 para revisões positivas e negativas), seguido pelo algoritmo RL-TF (0,84 para positivas e 0,85 para negativas). Já o algoritmo NB-TF apresentou o pior desempenho (0,80 para revisões positivas e 0,81 para negativas).

O algoritmo RL-TP obteve os melhores resultados nas métricas de precisão (0,86 nas revisões positivas e 0,85 nas negativas) e revocação (0,85 nas revisões positivas e 0,86 nas negativas). Enquanto o algoritmo NB-TF IDF obteve os piores resultados nas métricas de precisão (0,83 nas revisões positivas e 0,79 nas negativas) e revocação (0,78 nas revisões positivas e 0,84 nas negativas).

Os algoritmos que obtiveram os melhores resultados em cada métrica por coleção são sumarizados no Quadro 10 mostrado a seguir.

Quadro 10. Melhores classificadores para cada métrica por coleção

#	**Revisões de hotéis balanceada (TripAdvisor)**	**Revisões de hotéis desbalanceada (TripAdvisor)**	**Revisões de filmes**
F1	SVM-TF IDF	SVM-TF IDF / RL-TP	RL-TP
Precisão	SVM-TF IDF	RL-TF IDF	RL-TP
Revocação	SVM-TF IDF	SVM-TP / RL-TP	RL-TP

Fonte: Própria

Os algoritmos que obtiveram os piores resultados em cada métrica por coleção são sumarizados no Quadro 11 mostrado a seguir.

Quadro 11. Piores classificadores para cada métrica por coleção

#	Revisões de hotéis balanceada (TripAdvisor)	Revisões de hotéis desbalanceada (TripAdvisor)	Revisões de filmes
F1	NB-TF IDF	NB-TF IDF	NB-TF
Precisão	NB-TF IDF	NB-TF IDF	NB-TF
Revocação	NB-TF IDF	NB-TF IDF	NB-TF

Fonte: Própria

As abordagens tf e tp causaram pouca ou nenhuma variação dos resultados entre os algoritmos. Podemos inferir a partir disso que as informações que os atributos carregam em cada abordagem são muito similares, por isso a variação que ocorre é pouco significativa.

Já na abordagem TF IDF os atributos carregam informações a respeito da frequência do termo na revisão, e de sua raridade na coleção. Por isso a variação entre os resultados obtidos pelos algoritmos nas diferentes coleções é mais notável.

O algoritmo NB apresentou os piores resultados, nos três domínios, quando comparados aos algoritmos RL e SVM. Já os algoritmos RL e SVM apresentaram os resultados mais consistentes em ambas as coleções de revisões, e em todas as abordagens utilizadas.

O algoritmo SVM-TF IDF apresentou os resultados mais consistentes em todas as coleções mesmo com a variação de abordagem atributo-valor. É importante ressaltar que ele obteve um bom desempenho mesmo em uma base de dados desbalanceada.

4 CONSIDERAÇÕES FINAIS E RECOMENDAÇÕES

Esse trabalho de pesquisa relatou a análise de algoritmos de aprendizado de máquina supervisionado na tarefa de classificação de revisões. Nas seções a seguir são realizadas suas considerações finais e recomendações para trabalhos futuros.

4.1 CONSIDERAÇÕES FINAIS

Neste trabalho foi estudado o problema de análise de sentimento como problema de classificação da polaridade. Para que o estudo fosse possível foi necessária a realização de uma revisão da literatura do estado da arte em classificação de polaridade, a busca e a escolha de coleções e os domínios de revisões utilizadas e a seleção dos algoritmos de aprendizado de máquina supervisionados para experimentação e, por fim, foram relatados os resultados alcançados.

Podemos afirmar que a hipótese da pesquisa foi demonstrada ser acertada. Como pode ser vistos nos resultados demonstrados nas Tabelas 1, 2 e 3 e no Quadro 10, os algoritmos SVM e RL não só apresentaram um desempenho consistente em diferentes domínios, como também foram considerados os melhores classificadores de dois em três domínios.

Portanto, com base nos resultados obtidos foi possível perceber que os algoritmos Regressão Logística e SVM apresentaram uma consistência maior que o algoritmo Naive Bayes para as coleções de revisão de filmes, hotéis desbalanceada e balanceada. A Regressão Logística foi o que apresentou menor variação dos resultados entre as diferentes coleções.

Os melhores resultados foram obtidos nas bases de dados em que as instâncias das classes positivas e negativas eram balanceadas.

4.2 RECOMENDAÇÕES

Durante o desenvolvimento deste trabalho foi possível notar alguns aspectos para serem abordados em trabalhos futuros, dentre os quais:

a) Aplicação de n-gramas na montagem de vetor de atributos;

b) Explorar técnicas mais elaboradas para seleção de atributos;

c) Aplicação de algoritmos não-supervisionados em relação aos supervisionados;

d) Classificação sobre múltiplas entidades;

e) Identificação de múltiplos autores.

REFERÊNCIAS

AAS, Kjersti; EIKVIL, Line. Text categorisation: A survey. 1999.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**: The Concepts and Technology behind Search. ACM Press Books. 2011.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python: analyzing text with the natural language toolkit**. " O'Reilly Media, Inc.", 2009.

CAMBRIA, Erik; WHITE, Bebo. Jumping NLP curves: A review of natural language processing research. **IEEE Computational intelligence magazine**, v. 9, n. 2, p. 48-57, 2014.

CHAOVALIT, Pimwadee; ZHOU, Lina. Movie review mining: A comparison between supervised and unsupervised classification approaches. In: **System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on**. IEEE, 2005. p. 112c-112c.

DAVE, Kushal; LAWRENCE, Steve; PENNOCK, David M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: **Proceedings of the 12th international conference on World Wide Web**. ACM, 2003. p. 519-528.

DE CASTRO LUNARDI, Alexandre; VITERBO, José; BERNARDINI, Flavia Cristina. Um Levantamento do Uso de Algoritmos de Aprendizado Supervisionado em Mineração de Opiniões.

FERREIRA, Raoni Simões. **A wikification prediction model based on the combination of latent, dyadic and monadic features**. 2016. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2016. doi:10.11606/T.55.2016.tde-29112016-164654. Acesso em: 2018-04-15

GO, Alec; BHAYANI, Richa; HUANG, Lei. Twitter sentiment classification using distant supervision. **CS224N Project Report**, Stanford, v. 1, n. 12, 2009.

GUPTA, Vishal; LEHAL, Gurpreet S. A survey of text mining techniques and applications. **Journal of emerging technologies in web intelligence**, v. 1, n. 1, p. 60-76, 2009.

HAN, Jiawei; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques, Elsevier, 2012.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. In: **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2004. p. 168-177.

LEE, Chi-Hoon. Learning to combine discriminative classifiers: confidence based. In: **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2010. p. 743-752.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, v. 5, n. 1, p. 1-167, 2012.

LIU, Bing. **Sentiment Analysis and Subjectivity**. Handbook of natural language processing, v. 2, p. 627-666, 2010.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**, Cambridge University Press. 1ª edição, 2009.

MARTINS, Claudia Aparecida et al. Uma Experiência em Mineraçao de Textos Utilizando Clustering Probabilıstico e Clustering Hierárquico. **Instituto de Ciências Matemáticas e de Computação. São Carlos: Universidade de São Paulo**, 2003.

ORTIGOSA, Alvaro; MARTÍN, José M.; CARRO, Rosa M. **Sentiment analysis in Facebook and its application to e-learning.** Computers in Human Behavior, v. 31, p. 527-541, 2014.

PANG, Bo; LEE, Lillian. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In**: Proceedings of the 42nd annual meeting on Association for Computational Linguistics**. Association for Computational Linguistics, 2004. p. 271.

PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. In: **Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10**. Association for Computational Linguistics, 2002. p. 79-86.

SADEGH, Mohammad; IBRAHIM, Roliana; OTHMAN, Zulaiha Ali. Opinion mining and sentiment analysis: A survey. **International Journal of Computers & Technology**, v. 2, n. 3, p. 171-178, 2012.

SILVA, Josiane Rodrigues da et al. Detecção de opiniões e análise de polaridade em documentos financeiros com múltiplas entidades. 2015.

TANG, Huifeng; TAN, Songbo; CHENG, Xueqi. A survey on sentiment detection of reviews. **Expert Systems with Applications**, v. 36, n. 7, p. 10760-10773, 2009.

WANG, Hongning; LU, Yue; ZHAI, Chengxiang. Latent aspect rating analysis on review text data: a rating regression approach. In: **Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining**. ACm, 2010. p. 783-792.

WHITELAW, Casey; GARG, Navendu; ARGAMON, Shlomo. Using appraisal groups for sentiment analysis. In: **Proceedings of the 14th ACM international conference on Information and knowledge management**. ACM, 2005. p. 625-631.

WITTEN, Ian H; FRANK, E. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2016.

ZHANG, Harry. The optimality of naive Bayes. **AA**, v. 1, n. 2, p. 3, 2004.