JOANA GABRIELA RIBEIRO DE SOUZA

ANÁLISE DE SENTIMENTO POR MEIO DE APRENDIZADO PROFUNDO APLICADO A AVALIAÇÕES DE HOTÉIS

Dissertação apresentada a Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S729a
2018

Souza, Joana Gabriela Ribeiro de, 1994-
Análise de sentimento por meio de aprendizado profundo aplicado a avaliações de hotéis / Joana Gabriela Ribeiro de Souza. – Viçosa, MG, 2018.
ix, 57 f. : il. (algumas color.) ; 29 cm.

Orientador: Alcione de Paiva Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 52-57.

1. Redes neurais (Computação). 2. Hotéis - Avaliação. 3. Emoções - Análise. 4. Análise linguística. 5. Aprendizado do computador. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. II. Título.

CDD 22. ed. 006.32

JOANA GABRIELA RIBEIRO DE SOUZA

ANÁLISE DE SENTIMENTO POR MEIO DE APRENDIZADO PROFUNDO APLICADO A AVALIAÇÕES DE HOTÉIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 04 de setembro de 2018.

Giovanni Ventorim Comarela
Giovanni Ventorim Comarela

Regina Maria Maciel Braga Villela

Alcione de Paiva Oliveira
(Orientador)

Dedico este trabalho aos meus pais Edilson e Edna que sempre estiveram presentes e me incentivando.

Agradecimentos

Agradeço ao meu professor e orientador Alcione de Paiva, pelos conhecimentos repassados a mim durante as aulas e reuniões, paciência e compreensão que foram essenciais para realização deste trabalho e para meu desenvolvimento como profissional.

Agradeço à agência de pesquisa e financiamento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro para realização deste trabalho com dedicação exclusiva. Ao Departamento de Informática da UFV, seus professores e colaboradores que contribuíram para minha formação.

Agradeço aos meus pais, Edilson e Edna, por toda a dedicação, apoio, preocupação e incentivo aos meus estudos por todos estes anos. Aos meus irmãos Daniel e Gustavo pelo apoio e momentos de descontração. A Letinha, minha segunda mãe, pela preocupação e apoio e a todos os demais familiares.

Aos meus amigos e também família do mestrado Guidson, Lili, Bárbara, Daniel, Vinício, Aly, Fábio e Roberta (também a galerinha do almoço e café, não dá pra citar todo mundo), aos meus amigos da Atlética das Ciências Exatas, em especial a galera da peteca, obrigado por todo apoio, amizade, companheirismo, ajuda e pelos momentos de descontração.

Agradeço a todos que, diretamente e indiretamente, contribuíram para que eu chegasse até aqui e para que este trabalho fosse desenvolvido.

Por fim, mas não menos importante, agradeço a Deus que esteve sempre presente em todos os momentos, por ser meu sustento, proteção e por todas as oportunidades concedidas.

Milagres acontecem quando a gente vai à luta.
O Teatro Mágico

Sumário

Lista de Figuras vi

Lista de Tabelas vii

Resumo viii

Abstract ix

1 Introdução 1

1.1 Problema . . . 2
1.2 Hipótese . . . 3
1.3 Objetivo . . . 3
1.3.1 Objetivos Específicos . . . 3
1.4 Estrutura da Dissertação . . . 4

2 A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews 7

2.1 Introduction . . . 7
2.2 Related Work . . . 9
2.3 The *Corpus* . . . 12
2.4 The Deep Learning Approach . . . 12
2.5 Results . . . 16
2.6 Conclusions . . . 19

3 Desenvolvimento de um *Corpus* de *Reviews* de Hotéis em Português Brasileiro 20

3.1 Introdução . . . 20
3.2 Trabalhos Relacionados . . . 22
3.3 Desenvolvimento do *Corpus* . . . 23

3.4 Análises e Normalizações do *Corpus* 25
3.5 Conclusão 30

4 Redes neurais convolucionais para análise de sentimentos em *reviews* de hotéis em Língua Portuguesa **31**
4.1 Introdução 31
4.2 Trabalhos Relacionados 34
4.3 Apresentação do *Corpus* utilizado 36
4.4 Método empregado 37
4.5 Resultados e Discussão 40
4.6 Conclusão 47

5 Conclusões Gerais e Trabalhos Futuros **49**
5.1 Trabalhos Futuros 50

Referências Bibliográficas **52**

Lista de Figuras

2.1 Model structure. Each rectangle in the figure denotes a layer of the neural network. 15

3.1 Passos do processo de criação e normalização do *corpus*. 25
3.2 Exemplos de mudanças na polaridade removendo as palavras sem e não. Onde as cores verde, vermelho e cinza significam as polaridades positiva, negativa e neutra respectivamente. 27
3.3 Gráfico com as palavras/*uni-grams* mais comuns. 29
3.4 Gráfico com a distribuição de *bi-grams* no *corpus*. 29

4.1 Estrutura da rede convolucional. Cada retângulo na figura representa uma camada da rede. 41
4.2 Avaliação com nota 1 atribuída pelo viajante, porém não possui conotação negativa para a maioria das pessoas. 47

Lista de Tabelas

2.1 Results for Corpus 1 17
2.2 Results for Corpus 2 17
2.3 Results for Corpus 3 17
2.4 Results for Corpus 4 18
2.5 Results for Corpus 5 18

3.1 Algumas ocorrências encontradas no *corpus* 27
3.2 A tabela apresenta a distribuição desequilibrada das avaliações entre as 5 classes 28

4.1 Resultados para o *Corpus* 1. 42
4.2 Resultados para o *Corpus* 2. 42
4.3 Resultados para o *Corpus* 3. 43
4.4 Resultados para o *Corpus* 4. 43
4.5 Resultados para o *Corpus* 5. 44
4.6 Resultados para o *Corpus* 6. 44
4.7 Resultados para o *Corpus* 7. 45
4.8 Classificação da amostra por voluntários conforme as cinco classes possíveis do TripAdvisor. 46
4.9 Classificação da amostra por voluntários mescladas em 3 classes. 46

Resumo

SOUZA, Joana Gabriela Ribeiro de, M.Sc., Universidade Federal de Viçosa, setembro de 2018. **Análise de Sentimento por meio de aprendizado profundo aplicado a avaliações de hotéis**. Orientador: Alcione de Paiva Oliveira.

Análise de Sentimentos é uma área ativa de pesquisa e tem apresentado resultados promissores. Existem na literatura diversas abordagens capazes de realizar diferentes tipos de classificação com boa precisão. Tivemos como objetivo principal deste trabalho o desenvolvimento de modelo que utilize redes neurais artificiais, mais especificamente modelos baseados nas arquiteturas relacionadas com o aprendizado profundo, para a tarefa de classificação de polaridade de avaliações de hotéis escritas em Língua Portuguesa. Dois modelos de redes neurais convolucionais foram desenvolvidos para realizar a tarefa de classificação de polaridades. O primeiro consistiu em uma rede convolucional que possuía 15 camadas alcançando precisão acima de 95% para as classificações das classes positiva e negativa isoladamente e mantendo o balanceamento entre as classes, e acima de 90%, 67% e 80% para as classes positiva, neutra e negativa, respectivamente. O segundo foi baseado no primeiro, porém com adequações de hiper-parâmetros, redução e realocação de camadas, utilização de um vetor representação de palavras maior (200 dimensões) e de um *corpus* cerca de 10 vezes maior em quantidade de *reviews* e com número de *tokens* por *review* também superior, com arquitetura com 10 camadas. Essas modificações na estrutura permitiram obter resultados ainda melhores do que os observados no primeiro modelo, na maioria dos casos. Diferente do primeiro modelo, neste segundo foi possível classificar as avaliações conforme as 5 classes do TripAdvisor e a partir de amostra recolhida do *corpus* classificada por voluntários, fizemos um comparativo entre os resultados alcançados utilizando nosso modelo e a classificação gerada pelas pessoas, apontando que apenas em um caso as pessoas obtiveram maior precisão do que o modelo desenvolvido.

Abstract

SOUZA, Joana Gabriela Ribeiro de, M.Sc., Universidade Federal de Viçosa, September, 2018. **Sentiment Analysis through deep learning applied to hotel's reviews**. Advisor: Alcione de Paiva Oliveira.

Sentiment Analysis is an active area of research and it has presented promising results. There are several approaches in the literature capable of performing different types of classification with good precision. This work aimed to develop an approach that uses artificial neural networks, more specifically models based on architectures related to the deep learning, for the task of polarity classification of hotel reviews written in Portuguese. Two convolutional neural network models were developed to perform the task of classifying polarities. The first consisted of a convolutional network that had 15 layers. It obtained interesting results reaching precision above 95% for the positive and negative classifications alone, maintaining the balance between classes and above 90%, 67% and 80% for the positive, neutral and negative classes, respectively. The second one was based on the first, but with adaptations of hyper-parameters, reduction and relocation of layers, use of a vector representing greater words (200 dimensions) and a *corpus* about 10 times greater in quantity of reviews and with number of tokens by review also superior, with architecture with 10 layers. These modifications in the structure allowed to obtain even better results than those observed in the first model, in most cases. Unlike the first model, in this second we were able to rate the ratings according to the 5 TripAdvisor classes. In addition, from the sample collected from the *corpus* classified by volunteers, we compared the results achieved using our model and the classification generated by the people. The comparison pointed out that in only one case, people obtained greater precision than the developed model.

Capítulo 1

Introdução

Saber a opinião de outras pessoas a respeito de produtos, serviços e pessoas públicas em geral tem se tornado um fator diferencial de competitividade para as empresas de acordo com Pang & Lee (2008). Com a Internet, a facilidade para adquirir informações a respeito de alguma coisa se tornou cada vez maior, conforme Pang & Lee (2008). Existem diversos blogs e sites especializados em fazer *reviews* a respeito de diversos assuntos. Vários sites de compras possuem um espaço dedicado a avaliação de produtos e/ou serviços por parte do cliente. As redes sociais também são lugares onde as pessoas costumam dar sua opinião, além de outros espaços. Essas facilidades auxiliam as pessoas a tomarem suas decisões a partir da experiência de outras. Saber a opinião das pessoas é algo muito importante também para a indústria, uma vez que, com a avaliação de seus clientes eles podem fazer direcionamentos de *marketing*, melhorar ou desenvolver novos produtos ou serviços, além de conhecer melhor o seu mercado, segundo Becker & Tumitan (2013).

Desde o seu surgimento, as redes sociais online, vêm se tornando um ambiente cada vez mais popular na Internet. Nesses ambientes, as pessoas deixam de ser apenas consumidoras de opinião/conteúdo e passam a gerá-la/o. Existem espaços próprios para que os usuários possam discutir assuntos específicos. O Facebook possui espaços como grupos onde é possível desde discutir sobre algum assunto a realizar vendas, além de páginas públicas, espaço para propagação de conteúdo por parte de marcas, personalidades e qualquer outro usuário. O Twitter, possui um aspecto interessante, dado que é um *microblog* onde as pessoas falam o que querem, essa rede social se tornou a preferida para serviços de notícias, famosos e personalidades do mundo afora conforme Jesus (2014). Existem ainda diversas redes sociais como Instagram, Linkedin, Google plus dentre outras. O Twitter é bastante utilizado como *corpus* na literatura para as mais diversas tarefas que envolvam Processamento de

Linguagem Natural (PLN) (Pak & Paroubek (2010), Petrović et al. (2010), Cieliebak et al. (2017) e Sanguinetti et al. (2018)). Além de sites especializados ou que permitem a avaliação de produtos e serviços, e possuem trabalhos utilizando sites como por exemplo o TripAdvisor (Valdivia et al. (2017a), Guzman & Maalej (2014) e Souza et al. (2018a)).

Assim, devido a geração em massa de informação textual na *Web* provinda de *microblogs*, mídias sociais e demais ambientes, e a heterogeneidade de forma, tanto no formato dos dados quanto da linguagem empregada, torna a tarefa de Análise de Sentimentos interessante e desafiadora, tornando-a uma área ativa e multidisciplinar.

1.1 Problema

A possibilidade de captar as opiniões do público em geral gera interesse tanto na comunidade científica, que é levada a muitos desafios em aberto, quanto no mundo dos negócios, devido à notável variedade de benefícios que podem ser alcançados, desde *marketing*, *business intelligence* à previsão financeira. Porém, minerar opiniões e sentimentos é uma tarefa difícil, uma vez que envolve uma compreensão profunda da maioria das regras explícita e implícita, regular e irregular, sintática e semântica apropriadas de uma língua conforme Erik Cambria (2015).

Devido a expansão da utilização da Web e redes sociais, para se informar a respeito de algo não é necessário perguntar a um conhecido, basta pesquisar na Internet e possivelmente haverá uma quantidade de informações suficientes para tirar suas próprias conclusões. Para a indústria, que anteriormente precisava realizar enquetes e pesquisas direcionadas ao lançamento ou aceitação de algum produto e/ou serviço, a utilização das mídias sociais (redes sociais, blogs e microblogs, *reviews*, fóruns e comentários) como forma de adquirir tais informações é de extrema importância e ainda pode gerar economia para o setor de acordo com Liu (2012).

Portanto, podemos considerar pelo menos três razões para realizar estudos nessa área. Em primeiro lugar, pode ser aplicada em nas mais diversas áreas multidisciplinares: na indústria (onde ela mais se desenvolveu), economia, política, ciências sociais, e da PLN. Em segundo lugar, oferece uma gama de problemas que não foram estudados anteriormente. Em terceiro lugar, a quantidade enorme de dados relacionados a opiniões gerados pelas mídias sociais. Dados estes, que conforme a literatura, serviram de base para os avanços recentes realizados na área segundo Liu (2012).

As abordagens existentes para a análise do sentimento dependem principal-

mente de partes do texto em que as opiniões são explicitamente expressas, tais como termos de polaridade, palavras de afeto e suas frequências de co-ocorrência. No entanto, opiniões e sentimentos são muitas vezes transmitidos implicitamente, o que tornam ineficazes abordagens baseadas puramente na sintaxe de textos conforme Erik Cambria (2015).

A análise de sentimentos em idiomas diferentes do inglês, ainda é uma área pouco explorada. Trabalhos relacionados a essa área, utilizaram corpus rotulados em algum idioma traduzidos para o inglês como em dos Santos et al. (2014). Este é um ramo importante, dado que a quantidade de pessoas que geram dados que podem servir como base para análise de sentimentos que utilizam um idioma diferente do inglês é enorme. Por isso, é uma tarefa desafiadora desenvolver ferramentas que possuam suporte multilíngue, além de *corpus* de sentimentos em línguas variadas.

Assim, o estudo e desenvolvimento de técnicas, métodos, e demais aplicações na área de mineração de opiniões é de interesse, tanto a indústria quanto do meio acadêmico, pelas oportunidades e possibilidades da área dá-se a necessidade de investigá-la.

1.2 Hipótese

É possível, utilizando técnicas de aprendizado profundo, identificar a polaridade de avaliação de hotéis escritas em português brasileiro com acurácia semelhante ou superior ao estado da arte em classificação de polaridade.

1.3 Objetivo

Criação de modelo que utilize redes neurais artificiais, mais especificamente modelos baseados nas arquiteturas relacionadas com o aprendizado profundo, para a tarefa de classificação da polaridade de avaliação de hotéis escritas em Língua Portuguesa, que apresentem um desempenho superior aos métodos de classificação generativos.

1.3.1 Objetivos Específicos

- Criação de um modelo de rede neural profunda para classificação de polaridade.
- Criação de um *corpus* de avaliações de hotéis escritas em português brasileiro.

- Criação de um modelo de rede neural convolucional para classificação de polaridade.

1.4 Estrutura da Dissertação

Esta dissertação foi elaborada em formato de coletânea de artigos que foram produzidos durante a pesquisa. O primeiro capítulo apresenta uma introdução geral do problema discutido nessa dissertação e apresenta a hipótese e os objetivos gerais e específicos da pesquisa. Os três capítulos seguintes apresentam três artigos, que representam todo trabalho desenvolvido nessa dissertação, as conclusões gerais da pesquisa como um todo e possíveis trabalhos futuros, e por último todas as referências bibliográficas utilizadas.

O Capitulo 2 apresenta o artigo "*A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews*", em sua versão estendida, publicado nos anais da *23rd International Conference on Natural Language & Information Systems* (NLDB 2018). O artigo apresenta uma abordagem utilizando Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN) para realizar Análise de Sentimentos em um *corpus* de *reviews* de hotéis escritas em língua portuguesa. A escolha por trabalhar com CNN se baseou no fato desse tipo de rede ser capaz de capturar relações de localidade, algo interessante quando se trata da tarefa de processar texto, onde a posição das palavras é algo importante e que interfere em seu sentido. Como a CNN não consegue lidar diretamente com as palavras que compõem a *review*, foi utilizado *Global Vectors for Word Representation* (GloVe) como método de representação das palavras que serviram de entrada para a CNN. Assim, foi desenvolvida uma rede que foi capaz de classificar as *reviews* nas três possíveis classes: positiva, neutra e negativa. O *corpus* foi subdividido em 5 com o intuito de testar diferentes balanceamentos entre a quantidade de *reviews* positivas, neutras e negativas, uma vez que, o número de *reviews* positivas era cerca de 3 vezes o de neutras e 13 vezes o de negativas. Foram obtidos precisão e *recall* acima ou semelhante, mesmo sem realizar muitos procedimentos de pré-processamento do texto, a não ser a retirada das *stopwords*. Com a quantidade de *reviews* balanceadas conseguiu-se atingir precisão acima de 95% para as classes positiva e negativa isoladamente e acima de 90%, 67% e 80% para as classes positiva, neutra e negativa, respectivamente, conseguindo uma precisão considerada boa para a classe neutra que é de difícil identificação como descrito na literatura.

O segundo artigo "*Development of a Brazilian Portuguese Hotel's Reviews*

Corpus", submetido e aceito no *13th International Conference on the Computational Processing of Portuguese* (PROPOR 2018), é apresentado no Capítulo 3. O artigo apresenta o processo de criação e normalização de um *corpus* constituído por um conjunto de 730.069 *reviews* de hotéis escritos em língua portuguesa. As *reviews* foram recolhidas do site TripAdvisor (tripadvisor.com) no período entre fevereiro e março de 2018, capturando todas as *reviews* registradas dos hotéis selecionados até o período da coleta. Foram gerados quatro arquivos que conforme descrito neste capítulo estão disponíveis para acesso da comunidade, são eles "dates" que contém as datas de publicação de cada *review*, "notes" que contém as notas dadas pelos usuários (variando entre 1 e 5), "titles" que é o título ou resumo da *review* e por fim "comments" que contém o texto informado pelo usuário. Os quatro arquivos contêm todos os dados coletados linha por linha sequencialmente, de forma que a i-ésima linha do arquivo de "dates" corresponda à data da *review* da i-ésima linha no arquivo de "comments" e assim por diante. Um pré-processamento foi realizado em todos esses arquivos retirando as *tags* de *HyperText Markup Language* (HTML), além disso fizemos outras normalizações no arquivo "comments" para facilitar o seu uso ao utilizar ferramentas de Processamento de Linguagem Natural (PLN), uma vez que, este *corpus* foi desenvolvido com o intuito de utilizá-lo para realizar Análise de Sentimentos e a disponibilização do *corpus* à comunidade. Além disso, foram fornecidos alguns dados como os *uni-grams* e *bi-grams* mais comuns e tamanho máximo, mínimo e médio das *reviews*.

O terceiro artigo, "Redes neurais convolucionais para análise de sentimentos em *reviews* de hotéis em Língua Portuguesa", ainda não submetido, aguarda o parecer da banca, e está presente no Capítulo 4. O artigo propõe uma reformulação da estrutura de CNN utilizada no primeiro artigo Souza et al. (2018a) utilizando como *corpus* a base de dados descrita no segundo capítulo Souza et al. (2018b). Novamente, foi utilizado o GloVe como método de representação de palavras. A arquitetura desenvolvida possui complexidade menor e mesmo assim superou em sua maioria os resultados obtidos no primeiro capítulo. Subdividiu-se o *corpus* em 7 no intuito de testar diferentes balanceamentos entre as classes, além de classificar as avaliações não apenas nas polaridades negativa, neutra e positiva, mas também conforme à própria classificação do TripAdvisor, em 5 classes. Obtive-se novamente precisão e *recall* satisfatórios sem a realização de muitos procedimentos de pré-processamento. Foi realizada uma amostragem do *corpus* e essa amostra foi classificada por um grupo de voluntários nas 5 classes do TripAdvisor (horrível, ruim, razoável, muito bom e excelente) e utilizada como comparativo para nosso modelo que obteve resultados melhores em praticamente todos os casos, apontando

que o modelo possui acurácia satisfatória para a tarefa para o qual foi proposto.

Por fim, o Capítulo 5 contém um resumo dos resultados alcançados e as conclusões obtidas durante a pesquisa. Neste capítulo também foram incluídos possíveis trabalhos futuros para continuação da pesquisa.

Capítulo 2

A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews

Sentiment Analysis is an active area of research and has presented promising results. There are several approaches for modeling that are capable of performing classifications with good accuracy. However, there is no approach that performs well in all contexts, and the nature of the corpus used can exert a great influence. This paper describes a research that presents a convolutional neural network approach to the Sentiment Analysis applied to hotel's reviews, and performs a comparison with models previously used on the same corpus.

2.1 Introduction

Sentiment Analysis can be considered as the computational study of feelings, opinions and emotions expressed in the form of text Liu (2010). In general, the objective is to evaluate a given text searching not for its meaning, but for emotions searching for the emotions embedded in the sentences. Therefore, a text is usually divided into sentences which their emotional polarities are calculated. The polarity of a sentence is usually its classification among one of the classes: positive, negative or neutral. Subsequently, a sum of these polarities can be made in order to classify the predominant feeling in a given text. Sentiment Analysis has become an important decision-making tool and several machine learning techniques have been applied with the objective of obtaining analyzes with a good degree of prediction. A model

that has commonly been used as a baseline for Sentiment Analysis is the bag of words model, commonly applied to a generative technique, such as Naive Bayes as Agarwal et al. (2016) and Dinu & Iuga (2012). Although it is a simple model it is computationally efficient, easy to use and produces, in many situations, good results according Agarwal et al. (2016) and Jurafsky & Martin (2017). Even though a good technique, word bag models and Naive Bayes, it does not produce the best results in all situations, and discriminative models, usually has better results in relation to accuracy according Jurafsky & Martin (2000). One of the discriminative techniques that has been most currently applied in the of Sentiment Analysis is the use of Artificial Neural Networks. The recent growth of the neural networks approach as a machine learning tool was mainly due to the development of hardware and learning algorithms that enabled the implementation of networks with multiple layers called deep learning Goldberg (2017).

Regardless of the technique used, it is fundamental that there are datasets that can be used to feed these machine learning algorithms. One of the possible sources for generating data set for Sentiment Analysis is the extraction of user reviews of products and services such as travel and hotel booking sites. People conduct reviews on a variety of Web sites, such as on social networks, forums, and specialized websites. There are some example of websites such as TripAdvisor (www.tripadvisor.com), Trivago (www.trivago.com) and Booking (www.booking.com), among others that receive ratings related to tourism services. This type of site usually has a reserved area so the traveler can evaluate the location and service they have enjoyed using a star selector where the evaluator usually selects between one and five stars and can make a more detailed comment. When looking at travelers' reviews of hotels it may be noted that the number of stars in a rating is not always significant in itself, once people always look for more details. Therefore, the text informed by the evaluator is important, so that users can opt or not for such an option. In addition, it serves as a feedback, so the hotel can look for ways to improve or confirm positive points of the place.

When observing a series of evaluations made by people of hotels in Rio de Janeiro (the corpus used in this work), it is possible to notice that several people evaluate the hotel with high scores, but their commentary presents several problems found in the establishment. This type of information can not be acquired if only the assessment made by the number of stars is considered. Thus, analyzing the content of the evaluation becomes an important activity for the task of identifying the opinions expressed by the guests.

This paper describes a research that presents a convolutional neural network

approach to the Sentiment Analysis Applied to Hotel's reviews written in Brazilian Portuguese for the city of Rio de Janeiro, and performs a comparison with models previously used on the same corpus. Reviews were taken from tripadvisor site. This chapter is organized as follows: the next section presents researches previously developed that are related to this work; Section 2.3 describes the corpus used; Section 2.4 describes the model created to perform the analysis; Section 2.5 presents the results obtained; and Section 2.6 presents the final conclusions.

2.2 Related Work

The related works presented here are linked to the context of Sentiment Analysis focused on the tourism area or Sentiment Analysis using the deep learning approach.

The BESAHOT System, developed by Kasper & Vela (2011) used a statistical model, n-grams, along with specific domain and sentiment dictionaries, both developed by the authors through an information extractor. The aim is to classify polarity ratings of hotels written in German. They used a database classified by experts in which the evaluations were divided into segments with their polarity (positive, negative and neutral). When performing the classification process they obtained values of accuracy equal to 66%, 54% and 67% using the statistical classifier, the information extractor and the two methods together. Later, they removed neutral evaluations and treated only the positive and negative classes, so performance was improved by obtaining F-measure equal to 81% using the hybrid model. In our approach, we did not use domain-specific dictionaries and sentiments or n-grams, and our data did not need to be sorted manually by using the scores given by travelers as polarity identifiers.

Sodanil (2016) presented a multilanguage model, with hotel and restaurant ratings in English and Thai, from TripAdvisor and Agoda websites respectively. In this work, she proposed a weighted scheme for the properties of the hotel (room, location, service, price and facilities) and fourteen more subproperties. From these weights a polarity classification was made using Naive Bayes, Support Vector Machines (SVM) and Decision Tree algorithms. The best accuracy obtained was 92.61% using SVM in the fourth property. The accuracy of each algorithm was calculated according to the property and language. The best results were in the English language, but the process of obtaining data for training requires huge effort. An expert had to manually separate and classify all 3,000 ratings (half of each language) for each rating according to their polarity by category. In our work, we used as a way of

defining the polarity of the evaluation, the number of circles (stars) that the traveler attributed to the hotel, reducing the need for a specialist for this task. Even so, we obtained interesting results.

There are approaches that use Twitter (www.twitter.com) as the data source such as the one proposed by Shimada et al. (2011). In this work, the researchers have previously done a keyword extraction and "basic queries" (restaurants, events and facilities) based on the tourism portal of the Iizuca city in Japan. These information served as the searching basis for tweets (messages written by Twitter users up to 140 characters) related to the city. They also had to create rules and manual evaluation to filter tweets not related to tourism or related to other cities. After this pre-processing, they used the emoji "musical note" and "orz" (the first is positive and the second negative in the Japanese context) as seeds for capturing tweets that had polarity. The obtained corpus contained 116 tweets, a number considered low. The Naive Bayes technique was applied to the corpus and the accuracy (number of correct outputs/number of tweets in the dataset) obtained was 89% being the classification used with only the positive and negative classes. In our approach, we considered both the negative, neutral and positive classes, and only the negative and positive classes, and obtained interesting results. Our corpus was more robust and the process of pre-processing was not so manual, even when dealing with different sources of evaluation selection.

Context is an important aspect when it comes to review analysis. Context helps to get more information and disambiguate the meaning of the terms used by the reviewer. Taking that into account Aciar (2010) presented a method that uses rules for the identification of preferences and context of hotel reviews taken from TripAdvisor. The author selected 100 reviews to conduct her experiment. First, text mining tools were used to divide reviews into a set of sentences. After that, text mining tools and rule induction were used to define the rules that would be used to identify context and preferences. After baseline tests and 50 other reviews that were also manually classified by specialists, F-measure values of 88% and 91% were obtained for rules that identify context and preferences respectively. Our work does not seek to identify the context, but the focus is to determine the polarity of the reviews, although this information can serve as basis for future work that improves the determination of polarity.

Naive Bayes classifier is a method used by several researchers to perform Sentiment Analysis as in Shimada et al. (2011) and Martins et al. (2017). Our work performs an analysis of the same dataset used by Martins et al. (2017), which allows us to carry out a performance comparison between the two researches. Sentiment

Analysis was carried out using as dataset the comments, in Brazilian Portuguese, made on the TripAdvisor site about hotels in Rio de Janeiro city. Firstly, a normalization in the corpus was made converting all the characters to lowercase letter, besides withdrawal of accent and punctuation. The tests used were lemmatization, polarity inversion (after the word "no" the subsequent words receives a reverse polarity) and Laplace smoothing. The best result obtained 74.48% of accuracy by removing the stop-words and performing the inversion of polarity, lemmatization and Laplace smoothing. However, the accuracy for the neutral and negative values was 38.24% and 56.11% respectively. Classifying only the positive and negative classes using the same number of evaluations for both classes, the precision increased considerably from 90.83% to 97.9% and from 56.11% to 78.69% in the positive and negative classes respectively. In our work, we did not remove accents, punctuation, did not perform lemmatization, and did not convert the words into lowercase, in order to take advantage of aspects that may influence the refinement of the task of Sentiment Analysis. Even though we did not perform this pre-processing, we obtained relevant results in relation to Martins et al. (2017).

The studies mentioned so far have dealt only with research carried out in the field of tourism. None of these papers have used a deep learning technique. However, the deep learning approach has gained strong growth within Sentiment Analysis field. Therefore, since the cutting edge works use this approach, it is worth mentioning some of the recent works using this technique.

Tang et al. (2014) that have developed a deep learning system for message-level Twitter sentiment classification. The system uses as input the concatenating the sentiment-specific word embedding features with the state-of-the-art hand-crafted features. This input feeds a neural network with hybrid loss function to learn sentiment-specific word embedding, which encodes the sentiment information of tweets in the continuous representation of words.

Hassan & Mahmood (2017) developed a neural language model called ConvLstm. The architecture of its model is based on Convolutional Neural Network (CNN). To obtain better results, the model has an LSTM recurrence layer as an alternative to the grouping layer in order to efficiently capture the long term dependency. The model is more compact in comparison with networks that only adopt convolutional layers. The model was validated with two sentiment databases: the Stanford Sentiment Treebank and the Stanford Large Movie Review Dataset. Relation to this work, the similarity is in the choice of using databases with short texts, but the dataset is not directed to the tourism domain.

Sharath & Tandon (2017) applied deep learning in a twitter corpus with the

objective to apprehend the influences of words for each topic. The topic is an expression of different sentiment about various scenarios. They have created their own Word Embeddings and then performed inference on these embeddings to learn more about a word with respect to all the topics being considered, and also the top n-influencing words for each topic. Afterwards, they used these embeddings to predict the sentiment of the tweet with respect to a given topic, and all other topics under discussion. The developed model had two LSTM layers and a dense layer with softmax activation to yield the classification. It obtained a score above 64% in the classification of class 3 and 5. The F-score of the results can still be improved, but there is intention of improvement to the model.

2.3 The *Corpus*

Details on the creation of the corpus were presented in Martins et al. (2017). It consists of a set of 69,075 reviews about hotels located at Rio de Janeiro city, written in Brazilian Portuguese. The reviews were taken from the TripAdvisor site, and the most recent review date is from October 31, 2016.

Before applying the Sentiment Analysis techniques over the data, we first performed a pre-processing in the corpus to correct some issues. Empty lines, repeated punctuation characters, parentheses, and meaningless characters were removed. Also, because the tool that extracted comments from the site truncated them into 60 words, it was necessary to remove unfinished sentences. After these normalization the corpus resulted in 3,298,395 words and 51,408 types. The next stage of pre-processing was the removal of the stop-words except for the word "não" (no), since it's relevant when one want to extract an opinion from a particular topic. A common technique is to invert the polarity of the words in a sentence that occurs after the word "no", although this is not always the appropriate method [Kasper & Vela (2011)]. After the removal of the stop-words, the corpus resulted in 2,158,008 words and 42,284 types.

2.4 The Deep Learning Approach

The deep learning approach (DL) to natural language processing (NLP) has become quite popular in the last decade [Goldberg (2017)]. According to Goldberg (2017), this was due to the ease of model development, advances in hardware for parallel computing, and the creation of learning algorithms for neural networks with a large

number of layers. This last feature, a large number of layers, is what distinguishes deep neural networks from other neural networks. Particularly in the case of NLP the popularity is also due to the adherence of the technique to a large number of problems such as automatic translation, document summarization, syntactic and semantic annotation of lexical items, speech recognition, text generation, and Sentiment Analysis. In addition, one of the main advantages of applying DL to NLP is that in many cases, it is not necessary to manually design features, a slow and difficult task that requires specialized knowledge, since the features are detected by the model at the learning phase.

However, in the case of NLP text analysis, in order to adopt the DL approach, it is necessary to represent the words in a numerical form so that it can serve as input to the neural network. The simplest way is to use for each word a vector of the size of the vocabulary that contains 1 in the position corresponding to the word and 0 in all the others. This vector, called a one-hot vector, due to its simplicity, is not able to encode contextual information and co-occurrence relations between words. One way of numerically capturing contextual information and word relationships is through the use of dense vectors that encode these relationships through their positioning in an n dimensional space. There are some techniques for coding words in the form of dense vectors, with the best known being Word2vec [Mikolov et al. (2013)] and GloVe (global vectors) [Pennington et al. (2014)]. As it is considered the state of the art in the representation of words in the form of vectors, GloVe was the methodology used to represent the words in our work. According to Pennington et al. (2014), this method brings together two other methodologies: global matrix factorization and local context window. The first one is based on the use of statistical information to design the vectors that represents the words using their global co-occurrence, but the context task can be considered insufficient. The second uses local context windows which assists in catching the context of words, but it does not take advantage of the overall statistics of the corpus.

Preliminary tests showed that 100-dimensional GloVe vectors were sufficient for our purposes. They were generated over the entire corpus, after the preprocessing phase, but before the removal of the stop-words. We used the entire corpus in order to capture as much statistical information as possible about the context of the words aiming to create a vector that was best representative.

Another preparation step was the division of the corpus into 2 parts, one with the score given by the travelers and the other with the comments. For simplification, instead of using the five possible classes (horrible, bad, reasonable, very good and excellent), we mapped this five classes into just three classes (negative, neutral and

positive) according to the polarities presented in Liu (2012). In our mapping we have grouped the "horrible" and "bad" classes as the negative class, the "reasonable" class became the neutral class and the "very good" and "excellent" classes became the positive class.

For the development of the neural network model for a performance of the Sentiment Analysis, the Keras[1] framework was used. Keras [Chollet et al. (2015)] is an open source neural network library written in Python and it was chosen due to its high level API, and because it was developed with a focus on enabling fast experimentation. Keras is capable of running on top of TensorFlow[2], CNTK[3], or Theano[4], using them as backend. According to its documentation, Keras assists in the development of deep learning models, providing high-level building blocks, so it is not necessary to perform low-level operations such as product tensor and convolutions. Nevertheless, it needs a specialized and optimized tensor manipulation library to do so, serving as the "backend mechanism". In this work we used TensorFlow as backend. It is an open-source symbolic tensor manipulation framework developed by Google. We picked TensorFlow as it is actively maintained, robust, and flexible enough to be able to develop models target for both CPU and GPU.

In the case of sentences or sequence of words, the preferred neural network architecture is the recurrent neural networks and their variations, such as LSTM (Long Short-Term Memory) [Hochreiter & Schmidhuber (1997)] and GRU (Gated Recurrent Units) [Cho et al. (2014)]. However, in the case of sentences with a known maximum size, convolutional networks can also be applied, due to the fact that they are able retain locality information. For these reasons, the convolutional networks were the type of networks chosen for the development of our model. After several tests the structure of the convolutional network that produced the best results presented the layers shown in Figure 2.1.

The first layer of the network is the layer that receives the embedded sentences. It should be sized so that it can receive the sentence with the maximum size multiplied by the size of the vector of each word. Thus, in our case, the embedding layer has to have the dimension $(MaxLen, VectorSize)$, where $MaxLen$ is the review maximum size (i.e. 80), and $VectorSize$ is the size of the GloVe vector of each word, which in this case was established 100. Tests with 200-size vectors were performed, but the results were not better and lead to a longer training time. The

[1] https://keras.io
[2] https://www.tensorflow.org/
[3] https://github.com/Microsoft/cntk
[4] https://github.com/Theano/Theano

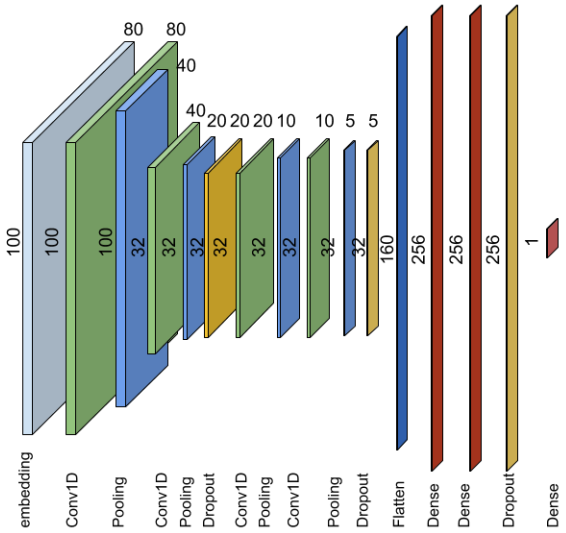

Figura 2.1: Model structure. Each rectangle in the figure denotes a layer of the neural network.

tests showed a structure with ten hidden layers, being four convolutional of one dimension, four pooling layers and two dense layers was able to produce good results without demanding high processing time. We also added three DropOut layer aiming to avoid overfitting, as suggest in Srivastava et al. (2014). For the pooling layers, we used as the basis of calculation the MaxPooling. At the output of each convolution layer we applied a Rectified Linear Unit (ReLU) activation function. The two dense layers emitted their results through a hyperbolic tangent activation function. The output layer had a single neuron using the sigmoid activation function, so the result would be a number between 0 and 1. For simplification purposes, we normalized the outputs to distribute them in each class according to the given interval: between 0 and 0.25 was assigned to the negative class which identification was 0. An output greater than 0.25 and up to 0.75 was attributed to the neutral class, identified as 1. It was attributed to the positive class, identified by 2, a result greater than 0.75.

We compiled the model with the Adam optimizer, and binary cross-entropy were used as the loss function. For training, we divided the corpus into two parts, 80% for training and 20% for testing. We performed several experiments in order to define the model parameters.

2.5 Results

The tests were carried out employing five different sub-corpus of the original corpus aiming to balance the numbers of reviews in each class: Corpus 1 (68078 reviews divided into 4863 negative, 12117 neutral and 52098 positive); Corpus 2 (29097 reviews divided into 4863 negative and 12117 neutral and positive); Corpus 3 (14589 reviews divided equally into each class); Corpus 4 (9726 reviews with the same amount of negatives and positives); and Corpus 5 (56960 reviews divided into 4863 negative and 52098 positive). All sub-corpus, except Corpus 1, had their reviews shuffled randomly to keep data from being concentrated, and the training and test split could capture examples from all classes involved. We did several tests to determine the number of training times for the model, and we identified that 50 epochs yield good results, and increasing the number of epochs did not result in greater accuracy, only a longer processing time.

Additionally, we tested word vectors (generated as GloVe vectors) of size 50, 100 and 200. The results presented here were obtained with the vector of size 100 as it achieved better results than vectors of size 50, and less processing time than those of size 200. On the other hand, when performing tests with vectors of size 200 we obtained results similar to those of size 100, but with a longer processing time. The results presented here correspond to the values obtained using a convolutional neural network architecture with the layout and data shape previously described.

The results obtained with our approach exceed those presented by Martins et al. (2017). Table 2.1 shows the precision and recall, and the confusion matrix generated using Corpus 1. Observing the confusion matrix, one can see that the greatest challenge is to identify the neutral class. The classes imbalance is one of the factors that influence this difficulty. Additionally, the neutral class ends up having a fuzzy boundary, which causes some difficulties. Another aspect that we can highlight is the normalization of the previous labels. The previous “bad” and “very good” classes ended up being classified as negative and positive respectively, but the neutral class has a broader range (between 0.25 and 0.75). The recall is relatively low for the negative and neutral classes, but the accuracy of our model was better than the work of Martins et al. (2017). The result shows an increase of 16.71% and 12.92% accuracy of the negative and neutral classes, respectively, and a result very similar to that found for the positive class accuracy (around 90%).

Table 2.2 shows the precision and recall found using Corpus 2. In this corpus the number of positive reviews has been reduced to the same number of neutrals. We can note that the performance for neutral reviews has improved considerably,

Tabela 2.1: Results for Corpus 1

	Precision and recall		Confusion matrix		
	Precision	*Recall*	Negative	Neutral	Positive
Negative	72.82%	56.78%	526	1380	79
Neutral	51.16%	56.19%	163	1013	913
Positive	90.47%	89.93%	33	1013	9384

and remained close to that found in the negative and positive classes compared to Corpus 1. With this balancing we obtained the best accuracy to identify the neutral class, above the results found in related works, pointing out that our approach was able to better classify the neutral class correctly, as can be observed in the confusion matrix presented in Table 2.2.

Tabela 2.2: Results for Corpus 2

	Precision and recall		Confusion matrix		
	Precision	Recall	Negative	Neutral	Positive
Negative	73.09%	61.85%	627	361	27
Neutral	73.36%	82.05%	215	1976	217
Positive	89.27%	84.19%	20	358	2016

Using Corpus 3, where all classes were balanced, we obtained the results presented in Table 2.3, indicating a higher hit rate for the positive and negative classes in relation to the other two corpus. However, in relation to the neutral class, there was an improvement over the results obtained in Corpus 1, but we lost precision in relation to Corpus 2. It is possible to observe that the results obtained in our work surpass the results obtained by Martins et al. (2017) in all classes, with improvement of 1.54%, 29.19% and 25.99% for the positive, neutral and negative classes, respectively. With this, we were able to demonstrate that our network is able to correctly classify most of the reviews, even the neutral class, which according to the literature is harder to identify and is often dismissed as having difficult identification.

Tabela 2.3: Results for Corpus 3

	Precision and recall		Confusion matrix		
	Precision	Recall	Negative	Neutral	Positive
Negative	82.1%	73.44%	710	246	10
Neutral	67.43%	79.78%	145	758	61
Positive	92.37%	86.63%	10	121	852

We also performed experiments using only negative and positive reviews, as Martins et al. (2017) did. Using Corpus 4, which contains the same number of reviews of each class, we obtained the results presented in Table 2.4. These results are notable as they show a good hit rate, and the obtained accuracy of the model was 95.74%. In relation to Martins et al. (2017) the precision for the positive class was 1.86% smaller but in relation to the negative class we had an expressive improvement of 16.78%. Table 2.4 also present the generated confusion matrix where it can be seen that the error rate obtained was relatively low.

Tabela 2.4: Results for Corpus 4

	Precision and recall		Confusion matrix	
	Precision	Recall	Negative	Positive
Negative	96.06%	95.47%	931	44
Positive	95.49%	96.05%	38	931

Finally, we used Corpus 5 which contain a high imbalance between the negative and positive classes. Looking at Table 2.5, it is noticed that even with the imbalance, we obtained a slightly higher precision for the negative class (from 78.69% to 88.94% in relation to the classification using the same number of reviews of each class performed by Martins et al. (2017)) and a precision of 98.31% for the positive class.

Tabela 2.5: Results for Corpus 5

	Precision and recall		Confusion matrix	
	Precision	Recall	Negative	Positive
Negative	88.94%	82.01%	806	180
Positive	98.31%	99.01%	102	10307

A challenging aspect of working with Sentiment Analysis is related to the fact that opinions are subjective. Evaluations made by users may present inconsistencies. For instance, in the corpus there are evaluations with positive scores while in the corresponding description the user mostly points out the negative aspects or suggestions to the place visited. This phenomenon can cause the classifier to label the review as negative or neutral according to the input text, but its class is positive due to the score given by the user. This type of occurrence is discussed in Valdivia et al. (2017b), where they performed a study that analyzed the polarity of the evaluations performed by people compared to the results obtained by three different methods of Sentiment Analysis.

2.6 Conclusions

The results obtained with the tests performed in different corpus setup showed that the convolutional neural network achieved considerably better results than the study used in the comparison. We also took advantage of the maximum information available in the corpus. Unlike Martins et al. (2017), Kasper & Vela (2011) and Sodanil (2016), we did not perform lemmatization, withdraw the punctuation marks, normalize the text in lowercase, classify the polarity of the previously words or manual classify the text. Also, using CNN eliminate the need for creating feature rules that can be difficult to develop and validate as in Aciar (2010). We used a much larger number of examples for training and testing in our model than Shimada et al. (2011) and also covered three reviews classes.

CNN is often applied to corpus that has as many as billions of tokens, but even for a corpus that can be considered relatively small, the results were satisfactory. Future works may use a larger corpus and perform the tests again to observe the accuracy of the approach for massive data sets.

Using CNN for sentence classification is an interesting approach when one is working with small sentences, but as the size of sentences grows, recurrent network should be a more adequate approach.

Capítulo 3

Desenvolvimento de um *Corpus* de *Reviews* de Hotéis em Português Brasileiro

O fornecimento de informação textual voluntária mediada pela Internet, e particularmente pela Web 2.0, proporcionou a oportunidade de criar grandes *corpora* linguísticos. Esses *corpora* podem servir como um recurso fundamental para o desenvolvimento de aplicativos focados em linguagem natural, especialmente aqueles que usam técnicas de aprendizagem profunda que exigem grandes conjuntos de dados. A Análise de Sentimentos é uma aplicação que se beneficia desses recursos disponíveis. Este artigo descreve a criação de um *corpus* destinado a suportar aplicações para Análise de Sentimentos. Consiste em avaliações de hotéis localizados nas capitais brasileiras e no Distrito Federal, escritos em português brasileiro. As avaliações que compõem o *corpus* foram retiradas do TripAdvisor e passaram por normalização e marcação de POS *tagging*. O objetivo principal é disponibilizá-lo para a comunidade para ser usado em tarefas de aprendizado de máquina voltadas para a linguagem natural.

3.1 Introdução

A Web permite que os usuários interajam, colaborem uns com os outros e compartilhem informações sobre diversos assuntos. Existem espaços digitais, como redes sociais, fóruns, sites de comércio eletrônico entre outros, onde as pessoas geralmente expressam suas opiniões. Nestes ambientes, os usuários costumam fazer uso de lin-

guagem informal, sendo muito comum o uso de gírias, abreviações, diferentes grafias de palavras existentes na gramática, além da criação de novos termos. Estes são alguns dos desafios encontrados pelos pesquisadores na compilação de *corpus* com dados extraídos da Internet.

Corpora são recursos-chave para o treinamento e teste de aplicações focadas no Processamento de Linguagem Natural (PLN). No entanto, a criação desses recursos pode ser bastante complexa e demorada. Definir a fonte dos dados e quais tipos de normalização serão feitas, são algumas das diversas tarefas a serem realizadas. Segundo Kilgarriff & Grefenstette (2003), a Web tem sido amplamente utilizada como *corpus*, devido à quantidade de dados disponíveis em vários idiomas e gêneros textuais, gratuitos e de fácil acesso. Patil (2017) apresentam as dificuldades e esforços envolvidos na extração de informações da web. Para extrair e armazenar dados textuais da Internet, que se tornarão úteis para ferramentas de PLN, é necessário a adoção de ferramentas de *web scraping* e a aplicação de várias etapas pré-processo aos dados para remover "ruídos" indesejáveis. Ocorrências como o uso de siglas, termos em diferentes idiomas, *emojis*, uso informal da língua entre outras situações são comuns em textos escritos por pessoas na Internet. Patil (2017) também aponta os desafios para a extração de textos na Web devido à variedade de conteúdos e formatos disponíveis que variam de um gênero para outro (de uma rede social para um site do governo, por exemplo). Meyer et al. (2003), listaram detalhes importantes a serem considerados ao definir o uso da Web como um *corpus*, como a seleção da ferramenta de pesquisa apropriada, o pré-processamento necessário, entre outros. Duran et al. (2014) e Hartmann et al. (2014) também apontam normalizações interessantes que foram levadas em consideração no desenvolvimento de um *corpus* composto por avaliações de produtos em português brasileiro.

Este trabalho foi motivado pela necessidade de um *corpus* em português brasileiro para a realização de Análise de Sentimentos. Para essa tarefa, foram escolhidas avaliações de hotéis para comporem o *corpus* considerando sua disponibilidade na Web. As informações textuais foram coletadas de um conhecido site de avaliações de atrações turísticas e passaram por várias etapas de processamento e ajustes no texto, que serão descritas neste capítulo. O *corpus* desenvolvido foi disponibilizado na Internet para ajudar a atender a necessidade de *corpus* em português brasileiro.

Este capítulo descreve os passos seguidos no desenvolvimento de um *corpus* de avaliações de hotéis em português brasileiro para a realização de Análise de Sentimentos. A Seção 3.2 apresenta alguns trabalhos relacionados. A Seção 3.3 descreve a compilação do *corpus* composto por textos escritos por pessoas que se hospedaram em hotéis nas 26 capitais brasileiras e no Distrito Federal. A Seção 3.4

apresenta o processo de análise e anotação dos tipos de ocorrências que originam termos fora do vocabulário neste tipo de texto, além dos procedimentos envolvidos no pré-processamento e normalizações e, finalmente, é apresentada breve conclusão juntamente com a indicação de trabalhos futuros na Seção 3.5.

3.2 Trabalhos Relacionados

A construção do *corpus* é uma das tarefas fundamentais da área de PLN. Aqui são apresentados alguns trabalhos relacionados ao desenvolvimento de *corpora* escritos na Língua Portuguesa.

Duran et al. (2014) aponta algumas questões sobre a padronização de um *corpus* de avaliação de produtos em português. Nesse *corpus*, os autores realizaram a rotulagem do papel semântico, Análise de Sentimentos, classificação e sumarização. Eles usaram o marcador MXPOST para a anotação de *part-of-speech* (POS) e usaram uma pequena porção do *corpus* para medir a precisão. Depois, criaram manualmente 4 *golden corpora* após a normalização da ortografia (incluindo palavras estrangeiras e entidades nomeadas); caso; pontuação; e uso de gírias da Internet. Eles observaram que as informações de caso obtiveram a maior taxa de correção, embora esperassem que as gírias tivessem um impacto maior (apenas 0,24% da amostra do *corpus*). Eles selecionaram quatro tarefas a serem executadas: (1) normalização de *true casing* usando Reconhecimento de Entidade Nomeada (REN) como uma das principais estratégias; (2) correção de problemas de pontuação; (3) correção ortográfica usando Unitex e, em seguida, verificação manual de palavras comuns para avaliar se a palavra foi corrigida com precisão e; (4) normalização de gírias da Internet, onde as palavras foram categorizadas como escritas diferentes de normas e abreviaturas cultivadas, repetição de sinais e letras e sequências relacionadas a emoções (como emoticons).

Hartmann et al. (2014) descrevem como o *corpus* usado em Duran et al. (2014) foi compilado e as estratégias usadas para normalizá-lo. Eles definiram 8 categorias com base dos tipos de ruído encontrados no *corpus*: erros de ortografia, sigla, nome próprio, abreviação, gíria da Internet, palavra estrangeira, unidade de medida e *tokens* não classificados ou duvidosos. Como resultado, eles criaram listas com os termos mais comuns para cada categoria e as concatenaram em uma ferramenta na qual é possível realizar a normalização de um texto (especialmente no contexto de análises de produtos). Na lista eles identificaram e corrigiram o acrônimo, nome próprio, abreviação, gíria da Internet, bem como a grafia de acordo com o verificador

ortográfico Aspell.

Bildhauer & Schäfer (2013) apresentou dois tipos de normalização (destrutiva e não-destrutiva) e uma arquitetura desenvolvida por eles para normalizar um *corpus* em alemão, sem perder informações que, para um *tagger* POS, podem ser considerados "ruídos" , mas podem dar pistas importantes sobre a linguagem não padronizada. A arquitetura foi baseada no uso de duas técnicas de normalização. Primeiro, uma normalização destrutiva foi realizada, usando HyDRA que une um padrão de regras com frequência de *n-grams* para definir quando uma palavra realmente contém um hífen, corrigindo-a. Subsequentemente, uma normalização não-destrutiva que visa manter o "ruído" do *corpus* reescrevendo da forma padrão palavras que foram enfaticamente escritas ("looooooove") ou com erros de digitação. Eles usaram uma camada de anotação para não perder informações que podem ser úteis para a Análise de Sentimentos ou para descobrir novos aspectos da linguagem, por exemplo.

Em Rocha & Santos (2000) é descrito como o *corpus* CETEM/Público foi criado (*Corpus of Extracts of Electronic Texts MCT/Public*). Um *corpus* desenvolvido com o apoio do Ministério da Ciência e Tecnologia (MCT). O CETEM foi criado a partir de textos jornalísticos do jornal "O público", fundado em 1990. O jornal é escrito em português, sendo quase exclusivamente de textos em português europeu, com exceções de textos escritos por brasileiros e africanos. As etapas de criação envolveram a limpeza de textos de imagens, classificação de assuntos e separação de sentenças, utilizando bibliotecas de programas desenvolvidas no projeto AC/DC. O resultado foi um *corpus* de 180 milhões de palavras. A principal diferença do presente trabalho é que o *corpus* não se destina à Análise de Sentimentos.

3.3 Desenvolvimento do *Corpus*

Como fonte de informações textuais para construir o *corpus* de avaliações de hotéis, foi escolhido o site TripAdvisor [1], pois é um dos sites mais utilizados pelos viajantes para avaliar não apenas hotéis, mas também vários outros tipos de atrações turísticas e serviços relacionados. Muitos estabelecimentos, principalmente relacionados com turismo, apresentam em sua recepção o selo de recomendação e/ou o certificado de qualidade atribuído pelo TripAdvisor. O *corpus* é composto apenas de avaliações de hotéis, por isso, sempre se falará sobre esse contexto de avaliação neste capítulo. No TripAdvisor, ao fazer uma revisão, os usuários devem inserir o

[1] https://www.tripadvisor.com.br

número de círculos (com significado semelhante ao de estrelas) correspondentes à avaliação geral do hotel, dar um título à avaliação (que pode ser entendida como um resumo), escrever a avaliação (com pelo menos 200 caracteres, onde se pode dar mais detalhes sobre a estadia), escolher o mês e ano da visita, bem como outras informações não obrigatórias. Foram coletados quatro informações das avaliações: o número de círculos, o título, a avaliação e, em vez da data da estadia, coletamos a data em que a avaliação foi realizada no site.

Os dados foram coletados apenas de acomodações classificadas pelo site como hotéis. As avaliações foram coletadas de fevereiro a março de 2018, portanto, a avaliação mais recente no *corpus* foi realizada em 20 de março de 2018. As avaliações foram feitas a respeito de hotéis das 26 capitais dos estados brasileiros e também no Distrito Federal. Optou-se por coletar avaliações de hotéis das capitais para ter um critério claro de delimitação do número de cidades e também para cobrir todos os estados brasileiros. Reunimos um total de 730.069 avaliações. Até o momento, o *corpus* não havia passado por nenhum pré-processamento linguístico, apenas a remoção das *tags* HTML. Usando o *Natural Language Toolkit* (NLTK) [2], uma biblioteca *Python* para PLN, verificou-se que o *corpus* continha 55.950.007 *tokens* e 457.337 tipos. Entre as palavras mais comuns estão:“hotel”, “não”, “bem”, “manhã”, “bom”, “quarto”, “café”, “localização”, “atendimento” e “excelente”. O que é esperado, dado o contexto das avaliações. Em contrapartida, os termos menos utilizados referem-se a erros de escrita, muito comuns nesse meio, devido a diversos fatores como os diferentes níveis de alfabetização dos usuários, erro de digitação, dentre outros.

A Figura 3.1 apresenta de forma resumida as etapas seguidas para o desenvolvimento do *corpus*. Em primeiro lugar, a captura das avaliações de hotéis que gerou quatro arquivos (*dates* (datas), *grades* (notas/círculos), *titles* (títulos) e *comments* (comentários/avaliações)) para cada hotel daquela capital. Essa captura foi possível a partir de coletor desenvolvido especificamente para essa coleta. Após este processo, foram reunidos todos os arquivos por cidade, por região e posteriormente em apenas quatro arquivos contendo todos os dados coletados linha por linha sequencialmente, de forma que a i-ésima linha do arquivo de *dates* corresponda à data da avaliação que está na i-ésima linha no arquivo de *comments* e assim por diante. Após essa junção executou-se a primeira normalização onde todas as *tags* HTML foram removidas e no caso das datas, convertidas do formato “2 de janeiro de 2018” para “02/01/2018” e para as notas converteu-se o código interno

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

usado pelo TripAdvisor de “bubble_10” até “bubble_50” para números de “1” a “5”. Posteriormente, a segunda normalização foi realizada, nela tentou-se fazer várias correções não-destrutivas conforme Bildhauer & Schäfer (2013) resultando em duas versões dos comentários: uma com as normalizações para *tokenização* descritas na seção posterior e outra também sem as *stopwords*. Estão sendo disponibilizados quatro arquivos (*comments normalized, dates, grades* e *titles*) com acesso livre para a comunidade [3].

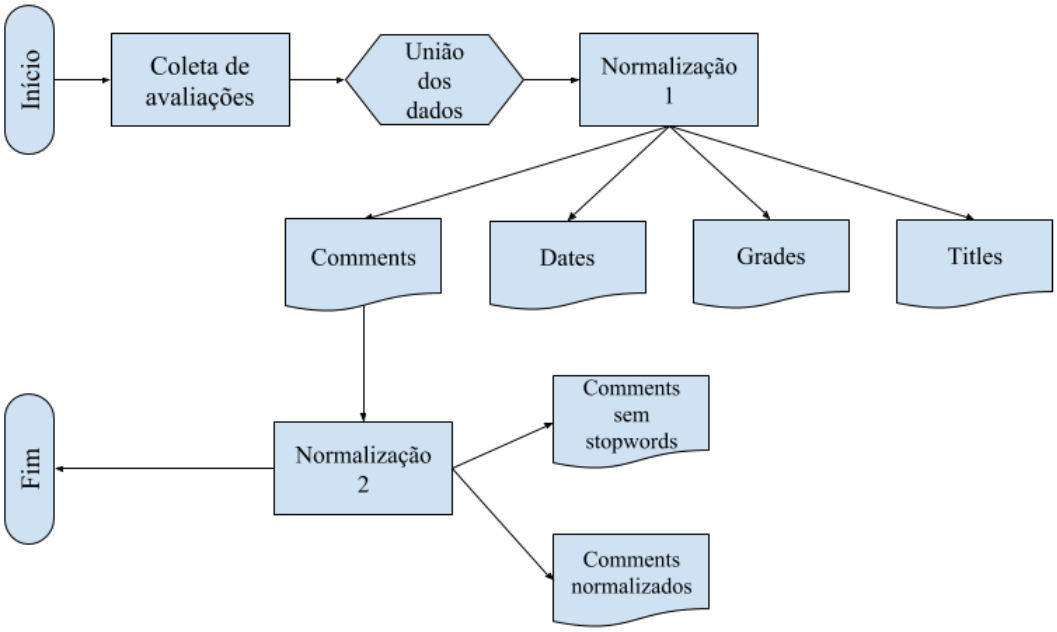

Figura 3.1: Passos do processo de criação e normalização do *corpus*.

3.4 Análises e Normalizações do *Corpus*

O NLTK foi utilizado para obter uma ideia geral do tamanho do *corpus* e com apoio dos trabalhos relacionados para estabelecer uma base de normalizações que pudessem e seriam interessantes para o uso que seria dado ao *corpus*. A primeira normalização no *corpus* foi a remoção dos excessos de pontuação e sequências de letras repetidas que não formavam palavras. Como o TripAdvisor exige que o usuário escreva um comentário de pelo menos 200 caracteres, em várias ocasiões os usuários completaram os comentários com sequências de pontuação e caracteres aleatórios que não formaram palavras, por isso removeu-se várias dessas ocorrências. Mantiveram-se reticências e sequências de até três exclamações ou pontos de

[3] Os arquivos do *corpus* estão disponíveis em: https://bit.ly/2JVRJbI

interrogação (que podem ter algum significado quando se trata da Análise de Sentimentos). Foi desenvolvido um dicionário léxico básico para ajudar a reduzir o número de palavras que estavam ligadas umas às outras (por exemplo, “Bomcafédamanhã”). Foram separados termos como números ou hifens (precedidos e seguidos por espaços) ligados à palavras (por exemplo, “8Limpeza” para “8 Limpeza” e “-Gostei” para “- Gostei”. Além disso, mantiveram-se as palavras da maneira como foram escritas, mesmo que incorretamente, devido a erros de digitação ou intenção do usuário. Com isso, termos como “adoreeeeei” e “lliiiixxxoooo” foram mantidos em sua forma original. Na Internet, é comum escrever termos maiúsculos como forma de enfatizar algo de forma positiva ou negativa, por essa razão também essas palavras foram mantidas com as letras maiúsculas ou minúsculas conforme digitado pelo usuário.

Observou-se que o *corpus* tinha vários tipos de erros na formatação das palavras que os impediam de serem *tokenizadas* corretamente, e esses erros não se encaixavam nos tipos de erros ou “ruídos” mencionados anteriormente. A Tabela 3.1 indica exemplos de ocorrências que eram muito comuns no *corpus* e sua correção tornou o *tokenizador* mais eficiente. Após essa normalização, o número de *tokens* aumentou (mesmo com a normalização destrutiva anterior) e o número de tipos foi consideravelmente reduzido, pois as correções dessas ocorrências permitiram que mais palavras puderam ser reconhecidas e contadas corretamente.

Fazendo alguns testes empíricos, notou-se que o *tokenizador* do NLTK falha em alguns casos comuns, como nos exemplos mostrados na Tabela 3.1. Também foi testado o *tokenizador* do Spacy [4], que é uma *Application Programming Interface* (API) também desenvolvida em *Python* para PLN, que de acordo com os desenvolvedores, é a ferramenta mais rápida e fornece o máximo de recursos atualizados. Além disso, o Spacy suporta aprendizado profundo, que é um tema em destaque na atualidade. No entanto, considerando os casos da Tabela 3.1, o Spacy não separa todos os *tokens*, apesar de separar alguns outros casos que o NLTK não consegue (“calçada..” tokenizado para [“calçada”, “.”, “.”] pelo Spacy e [“calçada..”] pelo NLTK), mas corrigidos boa parte desses problemas com o processo de normalização. A razão pela qual o NLTK foi selecionado como a ferramenta de PLN neste trabalho é devido a sua forma de tratar palavras que contêm hifens. O NLTK mantém o termo como um *token* único (“wi-fi” *tokenizado* para [“ wi-fi”]) enquanto o Spacy o trata como *tokens* distintos (“wi-fi” *tokenizado* para [“wi”, “-”,“fi”]). Como um dos objetivos do trabalho é disponibilizar o *corpus* para a comunidade, optou-se por manter esses

[4] https://spacy.io

termos assim. Após escolhida a ferramenta e realizadas as normalizações descritas acima, obteve-se uma contagem de 56.743.114 *tokens* e 246.307 tipos.

Considerando as *stopwords* e sinais de pontuação, notou-se que cada avaliação, em média, consiste em cerca de 77 *tokens*, com a maior avaliação tendo 2.857 *tokens* e a menor tendo apenas 2 *tokens*.

Tabela 3.1: Algumas ocorrências encontradas no *corpus*

Ocorrência	Correção
muito.Porém	muito. Porém
residência..	residência...
*apartamentos	* apartamentos
custo/benefício	custo / benefício
2km	2 km

Como o objetivo principal do *corpus* é ser usado em aplicações de Análise de Sentimentos, palavras como "não" e "sem" podem alterar o significado da frase invertendo sua polaridade, por exemplo (em Análise de Sentimentos, a polaridade de um termo é basicamente sua classificação entre as classes: positiva, negativa ou neutra). A Figura 3.2 contém exemplos de frases encontradas no *corpus* que ao remover essas palavras têm sua polaridade alterada. Devido a esse tipo de evento, *stopwords* que poderiam indicar mudança de polaridade, intensidade ou pistas para a próxima classificação de sentença, foram mantidas no *corpus* normalizado.

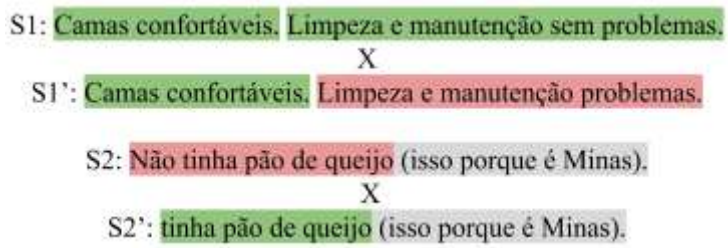

Figura 3.2: Exemplos de mudanças na polaridade removendo as palavras sem e não. Onde as cores verde, vermelho e cinza significam as polaridades positiva, negativa e neutra respectivamente.

Assim, ao manipular o conjunto de *stopwords* em português presentes no NLTK manteve-se os termos "não", "mais", "mas", "muito", "sem" e "nem". Após essa normalização, criou-se um *corpus* de avaliações sem *stopwords* com 39.165.169 *tokens*. Foi produzido um conjunto de quatro arquivos que podem ser usados pela comunidade para várias finalidades (não é fornecido o arquivo sem *stopwords*, pois

não foram removidas todas elas do *corpus*), além de servir como um recurso para pesquisa futura em Análise de Sentimentos.

Após essas normalizações, algumas análises foram feitas sobre o conteúdo do *corpus*. As avaliações são divididas em cinco classes: "horrível", "ruim", "razoável", "muito bom" e "excelente". A Tabela 3.2 apresenta a distribuição das avaliações nas cinco classes disponíveis, mostrando que há um desequilíbrio considerável entre as classes negativa, neutra e positiva, enquanto as classes negativas (horrível e ruim) juntas correspondem a apenas 7,1% do *corpus*, 16,6% das avaliações são neutras (razoável) e 76,2% são positivas (muito bom e excelente). Dependendo da aplicação, seria interessante equilibrar essas classes ou fazer algum tipo de compensação no algoritmo de aprendizado de máquina.

Tabela 3.2: A tabela apresenta a distribuição desequilibrada das avaliações entre as 5 classes

Distribuição de classes	
Classe	Porcentagem
Horrível	2,8%
Ruim	4,3%
Razoável	16,6%
Muito bom	40,2%
Excelente	36,0%

Embora tenham sido selecionados apenas comentários em português no site do TripAdvisor, o *corpus* resultante ainda contém *emojis* e muitas palavras em outros idiomas, principalmente termos em inglês e espanhol, até mesmo em chinês. Várias dessas avaliações datam da Copa do Mundo de 2014, que atraiu turistas de todo o mundo para as cidades sede. Ressalta-se que várias dessas ocorrências misturam o português com outras línguas, e optou-se por manter esses termos como foram escritos, sem traduções.

A Figura 3.3 exibe um gráfico com as palavras mais comuns do *corpus*, após eliminar os sinais de pontuação. Não surpreendentemente, preposições estão entre as palavras mais comuns. Além disso, outras palavras frequentes são altamente relacionadas ao contexto de hotéis, tais como: "hotel", "café", "quarto", "manhã", "serviço" e "localização". A Figura 3.4 mostra os *bi-grams* mais comuns no *corpus*, apontando que termos como "café da" e "da manhã" são os termos mais comuns, palavras que se referem diretamente ao contexto das avaliações.

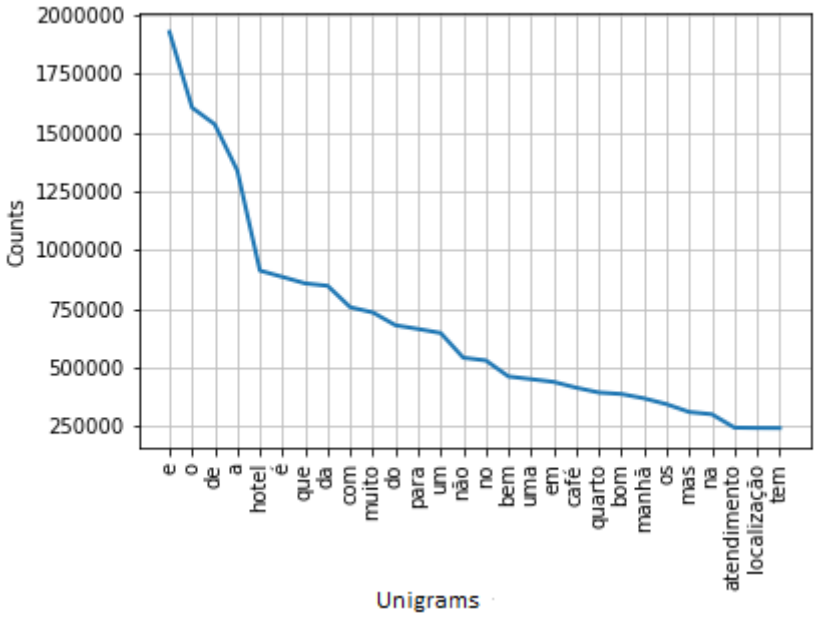

Figura 3.3: Gráfico com as palavras/*uni-grams* mais comuns.

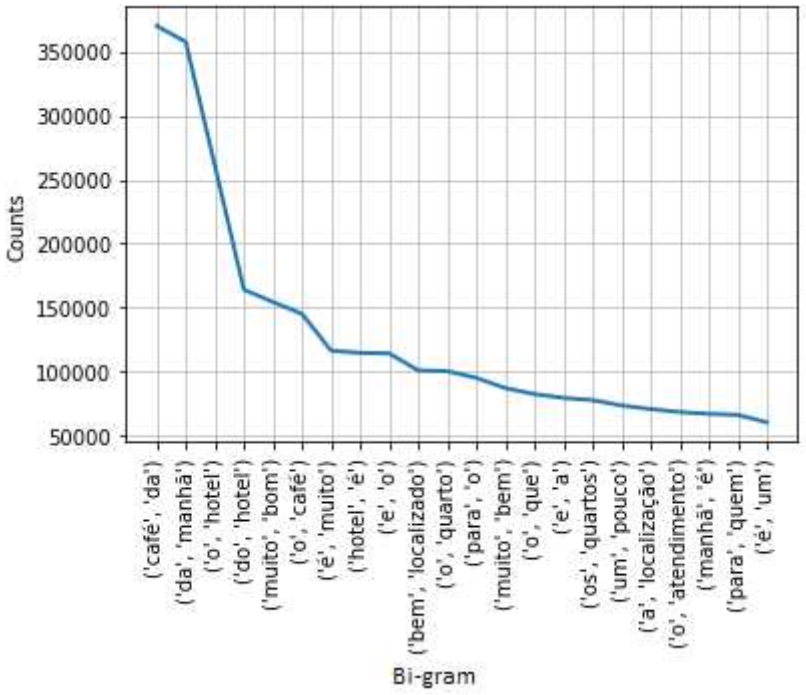

Figura 3.4: Gráfico com a distribuição de *bi-grams* no *corpus*.

3.5 Conclusão

A principal contribuição deste trabalho é a produção de um *corpus* de avaliações de hotéis com tamanho considerável que servirá como um conjunto de dados para trabalhos futuros em Análise de Sentimentos. Disponibilizando-o gratuitamente para a comunidade. Foi realizada uma normalização semiautomática para reduzir o "ruído" presente no *corpus*, mas com o intuito de mantê-lo o mais fiel possível, considerando que se trata de um *corpus* composto por textos extraídos da Web. O *corpus* também pode ser usado para auxiliar nas tarefas de extração de informações, identificação de padrões e novos aspectos presentes no contexto de avaliações de hotéis, entre outros. Como trabalhos futuros, técnicas como as de Bildhauer & Schäfer (2013) podem ser aplicadas para corrigir a ortografia dos termos "ruidosos" mantendo seu significado original, bem como normalizar o arquivo "*titles*". Além disso, é possível adotar o uso de metodologias multilíngues para tratar casos de avaliações escritas parcial ou integralmente em outras línguas diferentes do Português.

Capítulo 4

Redes neurais convolucionais para análise de sentimentos em *reviews* de hotéis em Língua Portuguesa

Análise de Sentimentos é uma área ativa de pesquisa e tem apresentado resultados promissores. Existem na literatura diversas abordagens capazes de realizar diferentes tipos de classificação com boa precisão. No entanto, não há uma abordagem que tenha bom desempenho em todos os contextos, e a natureza do *corpus* usado pode exercer uma grande influência nos resultados obtidos. Este artigo descreve uma pesquisa que apresenta uma abordagem utilizando uma rede neural convolucional para a tarefa de Análise de Sentimentos aplicada a avaliações de hotéis, e realiza comparações com um modelo previamente executado em um *corpus* menor e de mesma natureza, e com uma classificação realizada por seres humanos em amostra do *corpus* utilizado nesse trabalho, demonstrando que a abordagem gerou resultados próximos e/ou melhores que trabalhos da literatura.

4.1 Introdução

Com a *Web* 2.0 as pessoas passaram a ter mais acesso à rede de computadores e utilizá-la para os mais diversos fins, gerando grandes volumes de dados nos mais variados formatos, inclusive na forma de textos em linguagem natural. Para lidar com essa imensa quantidade de informação, uma vez que é muito oneroso tratá-la manualmente, temos a área da Linguística Computacional, que busca desenvolver ferramentas que processem eficientemente a linguagem natural em dispositivos com-

putacionais [OTHERO & MENUZZI (2005)]. Análise de Sentimentos é uma subárea da Linguística Computacional que pode ser considerada como o estudo computacional de sentimentos, opiniões e emoções expressas em forma de texto [Liu (2010)]. De uma forma geral, o objetivo é avaliar um dado texto e buscar nele não o significado de seu conteúdo, mas sim as emoções predominantemente apresentadas. Nesse sentido, conforme de Oliveira Gonçalves (2015) podem ser observados aspectos como: a polaridade que representa o grau de positividade, neutralidade ou negatividade de um termo; a subjetividade que se relaciona a métodos focados na classificação da subjetividade do texto (por exemplo textos informais são mais subjetivos que formais); a força ou intensidade que representa a intensidade de uma emoção, sentimento ou de uma polaridade; a emoção que indica o sentimento ou humor que o autor possui relacionado a algum assunto de acordo com Liu (2010); a opinião que representa um ponto de vista pessoal do autor em relação a algum assunto conforme Tsytsarau & Palpanas (2012).

A partir da popularização do uso de sites que provém serviço de *microblogging* como Twitter[1] e redes sociais como o Facebook[2], por exemplo, as pessoas foram capazes de expressarem sua opinião de forma direta na *Web* gerando conteúdo que serve de base para se trabalhar com Análise de Sentimentos como o *corpus* apresentado por Pak & Paroubek (2010). Diversas são as técnicas empregadas para a identificação da polaridade dos documentos. É comum na literatura a utilização do modelo de representação "*bag-of-words*", onde não são consideradas regras gramaticais e/ou ordem das palavras na(s) sentença(s), mas se mantém diversidade lexical, em conjunto com técnicas generativas como *Naive Bayes*, como acontece em Martins et al. (2017). Abordagem essa que apresenta bons resultados em diversas situações como citam Jurafsky & Martin (2017). Porém, abordagens que utilizem métodos discriminativos podem gerar resultados melhores em determinadas situações, conforme Jurafsky & Martin (2000). Uma das técnicas discriminativas que mais vêm sendo utilizadas são as Redes Neurais Artificiais (RNA). Além de abordagens híbridas que utilizam recursos léxicos como o SentiWordNet em suas versões 1.0 Esuli & Sebastiani (2007) e 3.0 Baccianella et al. (2010).

A falta de padronização e formatação dos dados retirados da *Web* é um dos aspectos desafiadores das pesquisas em Análise de Sentimentos de acordo com Ain et al. (2017). A falta de dados rotulados é um dos desafios enfrentados pela área, por essa razão abordagens que utilizam modelos de *Deep Learning* (DL) vem sendo adotadas pela sua capacidade de aprendizado que diminui sensivelmente a neces-

[1] https://twitter.com
[2] https://www.facebook.com

sidade de uma etapa prévia de engenharia de *features* [Goldberg (2017)], e sendo capazes de fornecer treinamento supervisionado e não-supervisionado de acordo com Vateekul & Koomsubha (2016). São utilizadas na literatura diversos tipos de redes neurais como Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN), Redes Neurais Recorrentes (*Recurent Neural Network* - RNN), *Long Short-Term Memory* (LSTM), dentre outras arquiteturas, conforme Zhang et al. (2018). Shirani-Mehr (2014) apresenta um comparativo entre diferentes arquiteturas de redes neurais (convolucionais, recursivas e recorrentes) em comparação ao classificador *Naive Bayes* para Análise de Sentimento em avaliações de filmes, mostrando que, utilizando DL, foram obtidos os melhores resultados.

A utilização da *Web* para a elaboração de *corpus* textuais com o propósito de alimentar sistemas de aprendizado de máquina é cada vez mais comum. Os trabalhos de Hartmann et al. (2014), Souza et al. (2018b), Rocha & Santos (2000) e Pak & Paroubek (2010) são alguns exemplos de *corpus* desenvolvidos para Análise de Sentimentos retirados da *Web*. A plataforma de *microblog Twitter* é a fonte mais comum para a criação desse tipo de *corpus* devido ao tamanho limitado de caracteres e ao fato de ser um canal de comunicação rápido, muito utilizado por pessoas e organizações para disseminar ideias, notícias, produtos, etc. Além dessa plataforma, existem também diversos sites mais específicos onde os usuários podem postar avaliações sobre produtos e serviços, juntamente com uma nota final sobre o item avaliado. Como exemplos de sites deste tipo podemos citar o TripAdvisor[3] que recebe avaliações itens ligados ao turismo, o Booking [4], ligado a avaliação de hotéis e sites de comércio eletrônico em geral, como a Amazon[5]. Estes sites possuem o beneficio adicional de incluírem uma classificação, feita por pessoas, associada ao texto postado. Neste trabalho optou-se por lidar com este tipo de dado onde já existe um certo nível de rotulação/polarização, por isso, escolheu-se utilizar o site TripAdvisor, sendo ele um dos sites mais utilizados por viajantes para avaliar/encontrar hotéis, voos, restaurantes e demais tipos de atrações turísticas, não só no Brasil, mas no mundo. Muitos estabelecimentos, principalmente relacionados com turismo, apresentam em sua recepção o selo de recomendação e/ou o certificado de qualidade atribuído pelo TripAdvisor. O processo de extração de informação dos sites e preparação do *corpus* é uma tarefa complexa, e esse processo está descrito em Souza et al. (2018b).

Esse capítulo é organizado da seguinte maneira: na Seção 4.2 apresentamos

[3]https://www.tripadvisor.com.br
[4]https://www.booking.com
[5]https://www.amazon.com

trabalhos encontrados na literatura relacionados ao problema; na Seção 4.3 apresentamos o *corpus* utilizado nesse trabalho e suas particularidades; na Seção 4.4 é descrita a abordagem que utilizamos para a tarefa de classificação das *reviews*; os resultados são apresentados na Seção 4.5; e por fim, as conclusões e trabalhos futuros estão contidos na Seção 4.6.

4.2 Trabalhos Relacionados

Apresentamos aqui alguns trabalhos relacionados encontrados na literatura que serviram de base para o desenvolvimento desta abordagem, seja como comparativo, seja como ajuda no desenvolvimento do método utilizado.

Yogatama et al. (2017) apresentam um comparativo entre modelos discriminativos e generativos para a tarefa de classificação de texto baseados em LSTM em termos de complexidade amostral e taxas de erros assintóticos. A partir de resultados obtidos empiricamente mostraram que para problemas onde há a introdução de novas classes na base de dados, modelos generativos respondem melhor do que modelos discriminativos, que podem sofrer do problema de *catastrophic forgetting* (esquecimento a respeito das classes aprendidas anteriormente) e maior sensibilidade ao *overfitting*, porém apresenta resultados melhores para bases de dados maiores, ao contrário dos modelos generativos que foram melhores para bases de dados menores. Como a base de dados utilizada neste trabalho pode ser considerada como grande (milhares de *tokens*), este trabalho provém maior suporte à escolha por optar em usar um modelo discriminativo.

Kim (2014) mostra que com uma rede neural com apenas uma camada de convolução é possível obter bons resultados para a classificação de texto, não apenas para Análise de Sentimentos, mas também outras tarefas. Para a camada de entrada utilizou tanto vetores treinados (*word2vec*) quanto vetores inicializados arbitrariamente e treinados na própria rede, uma camada de convolução com três filtros, *DropOut* para evitar *overfitting* e a camada de saída utilizando *softmax*. Este trabalho apresenta a construção de uma estrutura de rede neural convolucional relativamente simples apontando o modo como foram configurados os hiper-parâmetros da redes, sendo utilizado como base para diversos trabalhos presentes na literatura.

dos Santos & Gatti (2014) desenvolveram um trabalho onde um dos aspectos mais interessantes em sua arquitetura de rede é a utilização de duas camadas convolucionais que permitem manipular palavras e sentenças de tamanho diversos. A rede contém como entrada um vetor de representação das palavras, criado com *word2vec*

ou arbitrariamente (mostrando resultados melhores com o vetor treinado) e posteriormente possui uma camada convolucional que retira *features* no nível de caracteres das palavras. Utilizaram essa rede para classificar um *corpus* de *tweets* e um de *reviews* de filmes. Esse experimento mostrou que a utilização da camada convolucional para *features* no nível de caractere é interessante para o *corpus* do *Twitter*, mas não muito relevante para o de *reviews* de filmes. Nesse capítulo utilizou-se um vetor pré-treinado pois uma camada de convolução no nível de caracteres talvez não seria interessante pois nos dados não existem tantas ocorrências de *hashtags* que são muito significativas, além de que realizaram a classificação no nível de sentença e neste capítulo é em nível de documento.

É muito comum em sites de *e-commerce* e outros, as pessoas terem a possibilidade de dar um número de estrelas para determinado produto/serviço, dessa forma é possível identificar rapidamente sua aprovação perante ao público. April & Daryl (2015) tiveram a ideia de a partir do texto de *reviews* gerar sua classificação em uma escala de 5 estrelas utilizando técnicas de DL. Utilizaram dois modelos de rede, uma LSTM e CNN, ambas com resultados ruins 51% e 34% de acurácia. A rede CNN foi baseada no modelo apresentado por Kim (2014), inclusive testaram a mesma configuração para a base de dados que utilizaram (*Yelp Open Dataset*[6], que para o experimento continha 1,6 milhões de *reviews*), porém a arquitetura de Kim (2014) é pensada para classificação binária e por sentença e o problema de April & Daryl (2015) era pra realizar uma classificação em 5 classes para a *review* como um todo. Semelhante a April & Daryl (2015) este trabalho visa classificar as *reviews* como um todo e em mais classes, não apenas em positivas e negativas.

Hassan & Mahmood (2017) desenvolveram um modelo de linguagem neural chamado ConvLstm. A arquitetura do modelo é baseada em CNN. Para obter melhores resultados, o modelo possui uma camada de recorrência LSTM como uma alternativa à camada de *pooling* com o intuito de reduzir a perda de informações locais detalhadas e capturar dependências de longo prazo na sequência de sentenças. O modelo é mais compacto em comparação com redes que adotam apenas camadas convolucionais. O modelo foi validado com dois bancos de dados de sentimentos: o *Stanford Sentiment Treebank* e o *Stanford Large Movie Review Dataset*.

Attia et al. (2018) apresenta um modelo de classificação multilinguagem e multi-classes utilizando uma rede neural em cinco camadas (*embedding*, *Conv1D*, *GlobalMaxPooling* e duas camadas totalmente conectadas), onde classificaram três bases de dados normalizadas para Análise de Sentimentos em diferentes idiomas

[6]https://www.yelp.com/dataset

(inglês, alemão e árabe). A camada de *embedding* inicializada arbitrariamente, sem utilização de *word embeddings*, transformando as palavras das sentenças em um *feature map* preservando informações espaciais. Utilizaram *ReLU* como função de ativação e *softmax* na camada de saída. Com essa arquitetura eles obtiveram resultados melhores ou próximos aos melhores resultados encontrados para tais bases. Diferente deles, optamos por utilizar *word embeddings*, uma vez que de qualquer forma era preciso codificar o texto para servir de entrada para a rede, além disso utilizaram *oversampling* para balancear as classes, algo que talvez não traga grandes benefícios ao modelo, apesar de terem conseguido pequenas melhoras na acurácia.

Souza et al. (2018a) apresentam uma abordagem utilizando uma CNN para realizar Análise de Sentimentos em *reviews* de hotéis escritas em português. Este trabalho realiza um comparativo com o trabalho de Martins et al. (2017) que utilizou o mesmo corpus, e obteve resultados gerais melhores, inclusive para a detecção da classe neutra, que conforme a literatura, é a de mais difícil detecção. Eles desenvolveram uma arquitetura utilizando uma camada de *embeding* com a entrada dos vetores das palavras codificadas utilizando o *GloVe* como *word embedding*, além das camadas de convolução, de *pooling*, de *DropOut*, camadas totalmente conectadas e da camada de saída com um neurônio usando a função de ativação *sigmoid*. A abordagem desenvolvida e apresentada nesse capítulo utilizou o trabalho de Souza et al. (2018a) como base, porém utilizando um outro *corpus*, maior, e modificando a estrutura da rede deixando-a menos complexa.

4.3 Apresentação do *Corpus* utilizado

O *corpus* utilizado neste trabalho é descrito em detalhes em Souza et al. (2018b). Ele consiste de 730.069 *reviews* de hotéis das capitais dos estados brasileiros e do Distrito Federal, escritas em língua portuguesa, retiradas do site TripAdvisor. As *reviews* datam de 27/05/2004 até 20/03/2018. Em Souza et al. (2018b) é realizado um processo de normalização e então são gerados 4 arquivos[7]: datas em que foram escritas as *reviews*, notas (de 1 a 5), títulos/resumo da *review* e por último a avaliação/comentário que são disponíveis para acesso.

Neste trabalho utilizou-se os arquivos dos comentários para a tarefa de Análise de Sentimentos e o de notas como *labels* para a definição da polaridade dos comentários. Os comentários e notas estão organizados linha por linha sequencialmente, de forma que a i-ésima linha do arquivo de notas corresponda à nota do comentário

[7]Que podem ser acessados em: https://bit.ly/2JVRJbI

da *review* que está na i-ésima linha no arquivo de comentários e assim por diante. Para melhor entendimento, será tratado aqui o texto escrito pelo viajante como "comentário" ou "avaliação" e o termo *"review"* como o todo: a data, a nota, o título e o comentário.

Neste trabalho foram feitos alguns ajustes, segmentações e experimentos tanto nas notas quanto nos comentários.

Considerou-se as notas de três formas, com as 5 classes (horrível, ruim, razoável, muito bom e excelente), com 3 classes (negativa, neutra e positiva) e com as classes negativa e positiva apenas. Para a representação das notas em 3 classes foram concatenadas as classes "horrível" e "ruim" em negativa, "razoável" em neutra e "muito bom" e "excelente" em positiva; já para 2 classes retirou-se a classe neutra e manteve-se as outras duas.

No arquivo de comentários, cada linha representa todo o texto da avaliação, podendo haver várias sentenças numa mesma linha, sendo que nessa abordagem o objetivo é identificar a polaridade do comentário todo. Utilizou-se tanto o *corpus* normalizado completo, quanto o sem *stopwords* ambos descritos em Souza et al. (2018b).

4.4 Método empregado

O uso de DL para PLN tornou-se bastante popular na última década [Goldberg (2017)]. Que segundo ele, isso se deveu à facilidade de desenvolvimento de modelos, avanços em *hardware* para computação paralela e à criação de algoritmos de aprendizado para redes neurais com um grande número de camadas. Possuir diversas camadas, é o que distingue redes neurais profundas (DL) de outras redes neurais. Particularmente no caso da PLN, a popularidade também se deve à aderência da técnica a um grande número de problemas, como tradução automática, sumarização de documentos, anotação sintática e semântica de itens lexicais, reconhecimento de fala, geração de texto e Análise de Sentimentos. Além disso, uma das principais vantagens da aplicação da DL a PLN é que, em muitos casos, não é necessário desenvolver *features* manualmente, tarefa lenta e difícil, que requer conhecimento especializado, uma vez que os recursos são detectados pelo modelo na fase de aprendizado.

No entanto, para a adoção de redes neurais para PLN, é necessário representar as palavras em uma forma numérica para que possam servir como entrada para a rede neural. A maneira mais simples é usar para cada palavra um vetor do ta-

manho do vocabulário que contém 1 na posição correspondente à palavra e 0 em todas as outras. Este tipo de codificação é chamada de *one-hot vector*, devido à sua simplicidade, ele não é capaz de codificar informações contextuais e relações de co-ocorrência entre palavras. Uma maneira de capturar numericamente informações contextuais e relações entre palavras é através do uso de vetores densos que codificam essas relações através de seu posicionamento em um espaço n dimensional. Existem algumas técnicas para codificar palavras na forma de vetores densos, os mais conhecidos são *Word2vec* [Mikolov et al. (2013)] e *GloVe* [Pennington et al. (2014)], que inclusive são usados em diversos trabalhos da literatura (por exemplo Souza et al. (2018a), Kim (2014) e dos Santos & Gatti (2014)). Por ser considerado o estado da arte na representação de palavras na forma de vetores, o *GloVe* foi escolhido como método para representação das palavras em neste trabalho. De acordo com Pennington et al. (2014), esse método é composto pela união de duas metodologias: fatorização da matriz global e janela de contexto local. O primeiro é baseado no uso de informações estatísticas para projetar os vetores que representam as palavras usando sua co-ocorrência global, mas a tarefa de identificação do contexto pode ser considerada insuficiente. O segundo usa janelas de contexto locais que auxiliam na captura do contexto das palavras, mas não aproveitam as estatísticas gerais do *corpus*. Como uma junção dessas duas estratégias, o *GloVe* se torna um método atrativo para a representação das palavras, pois já carrega informações que servirão de *features* para a rede neural.

Foram realizados testes que apontaram que vetores gerados no *GloVe* com 200 dimensões foram suficientes para o propósito deste trabalho. Então, foi gerado um arquivo que continha as representações das palavras, a partir do *corpus* de comentários como um todo, contendo as *stopwords*, sinais de pontuação além de não normalizar as palavras para letra minúscula. Todo o corpus foi utilizado para capturar o máximo possível de informações estatísticas sobre o contexto das palavras, com o objetivo de criar vetores mais representativos.

Para o desenvolvimento deste modelo de rede neural utilizou-se o *framework* Keras [8]. O Keras é uma biblioteca de rede neural de código aberto escrita em *Python*, foi escolhida por ser uma API (*Application Programming Interface*) de alto nível, e porque foi desenvolvida com foco em permitir a experimentação rápida. O Keras é capaz de rodar utilizando TensorFlow[9], o CNTK[10] ou o Theano[11], usando-os

[8] https://keras.io
[9] https://www.tensorflow.org/
[10] https://github.com/Microsoft/cntk
[11] https://github.com/Theano/Theano

como *backend*. De acordo com sua documentação, o Keras auxilia no desenvolvimento de modelos de aprendizagem profunda, fornecendo blocos de construção de alto nível, por isso não é necessário realizar operações de baixo nível. No entanto, ele precisa de uma biblioteca especializada e otimizada de manipulação de tensores para fazê-lo, servindo como o mecanismo *backend*. Neste trabalho para tal usamos o TensorFlow. Ele é uma estrutura de manipulação de tensores simbólicos de código aberto desenvolvida pelo Google. O TensorFlow foi escolhido por ser mantido ativamente, por ser robusto e flexível o suficiente para ser capaz de desenvolver modelos de destino para CPU e GPU.

No caso de sentenças ou sequência de palavras, a arquitetura de rede neural preferida são as redes neurais recorrentes e suas variações, como LSTM Hochreiter & Schmidhuber (1997) e GRU (*Gated Recurrent Unit*) Cho et al. (2014). No entanto, no caso de sentenças com um tamanho máximo conhecido, redes convolucionais também podem ser aplicadas, devido ao fato de que elas são capazes de reter informações de localidade. Por estas razões, as redes convolucionais foram o tipo de redes escolhidas para o desenvolvimento deste modelo. Chegou-se a implementar uma arquitetura semelhante a de Hassan & Mahmood (2017) utilizando uma camada LSTM como camada de *pooling*, porém não foram obtidos bons resultados (entre 50% a 51% de precisão para classes positivas e negativas), acredita-se que um dos motivos possa ter sido o tamanho dos comentários, pois devido o seu tamanho, 600 tokens, a camada LSTM tenta capturar muitas informações o que acaba não ajudando na generalização do modelo e deixando o modelo mais lento (cerca de 50% de tempo a mais para cada época de treinamento), por essa razão optou-se por utilizar camadas de *Maxpooling* para reduzir a dimensionalidade e abstrair as *features*.

O modelo construído foi uma simplificação da rede neural profunda apresentada em Souza et al. (2018a). Para definir os hiper-parâmetros da rede, e outros aspectos relacionados à criação do modelo utilizou-se também Zhang & Wallace (2015) como base. A Figura 4.1 apresenta o modelo desenvolvido que obteve melhores resultados. A camada de *embedding* é a primeira camada da rede que recebe as avaliações codificadas como uma matriz de tamanho: tamanho da avaliação x tamanho do vetor de palavras. Para o treinamento e teste utilizou-se o *corpus* sem *stopwords*. Utilizaríamos o tamanho da avaliação como o tamanho da maior avaliação (1834 *tokens*), porém isso gera grande tempo de processamento e muitas matrizes com zeros pois ao observar o *corpus* checou-se que haviam apenas 220 avaliações com mais de 600 *tokens*, por entender que uma quantidade relativamente baixa de avaliações seriam cortadas, optou-se por aderir a esse limite de entrada. O

tamanho de cada palavra é o tamanho do vetor gerado utilizando do *GloVe* que foi definido como 200 devido a testes realizados, onde ao aumentar o tamanho do vetor não houve melhora na acurácia, apenas maior tempo de processamento. O modelo possui ainda duas camadas de convolucionais intercaladas por 2 camadas de *pooling* e 3 camadas totalmente conectadas. Com essa estrutura obteve-se resultados bons montando uma rede menor e que recebe avaliações muito maiores do que as classificadas em Souza et al. (2018a). Adicionou-se também 2 camadas de *DropOut* no intuito de reduzir o *overfitting*, conforme sugerido em Srivastava et al. (2014). Utilizou-se nas camadas de *pooling* a função de *MaxPooling*. Nas camadas convolucionais foi selecionada a função de ativação *ReLU* (*Rectified Linear Unit*) que é uma das mais utilizadas para redes convolucionais como sugerido na literatura [Zhang & Wallace (2015)]. Nas camadas "densas" ou totalmente conectadas, utilizou-se *ReLU* como função de ativação nas camadas mais internas e na de saída *softmax* com o número de neurônios em função da quantidade de classes que se queria classificar. Como foram feitas algumas subdivisões no *corpus*, foram realizadas classificações em 2 classes: negativa e positiva, 3 classes: negativa, neutra e positiva e em 5 classes: "horrível", "ruim", "razoável", "muito bom" e "excelente", tendo um neurônio para cada classe. Na Figura 4.1, a camada de saída possui 3 variações (2-3-5) indicando a quantidade de classes que serão classificadas.

O modelo foi compilado utilizando o otimizador *Adam* e a função de perda *categorical_crossentropy*. Portanto, os arquivos de notas foram transformados, apresentados anteriormente, no formato de *one-hot vector* para utilização desta função de perda. Para o treinamento, dividiu-se o *corpus* em duas partes, 80% para treinamento e 20% para teste. Foram feitos diversos experimentos para definir os melhores hiper-parâmetros para o modelo baseados na literatura e nos resultados obtidos.

4.5 Resultados e Discussão

Os testes foram realizados com sete diferentes partições do *corpus* original, com o objetivo de equilibrar o número de avaliações nas classes: *Corpus* 1: 730069 avaliações divididas em 52000 negativas (7,1%), 121235 neutras (16,6%) e 556834 positivas (76,2%); *Corpus* 2: 294470 avaliações divididas em 52000 negativas (17,66%), 121235 de neutras (41,17%) e de positivas (41,17%)); *Corpus* 3: 156000 avaliações divididas igualmente em cada classe; *Corpus* 4: 104000 revisões divididas igualmente entre as classes negativa e positiva ;*Corpus* 5: 608834 avaliações divididas em 52000 negativas (8,54%) e 556834 positivas (91,46%); *Corpus* 6: 730069 avaliações divididas

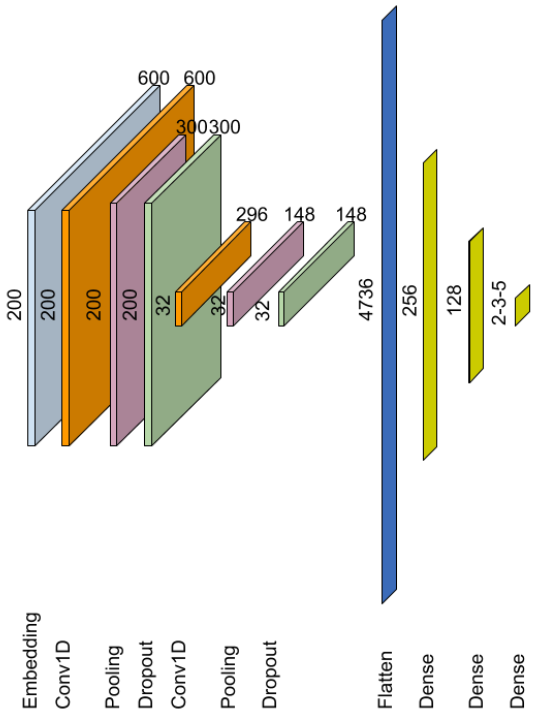

Figura 4.1: Estrutura da rede convolucional. Cada retângulo na figura representa uma camada da rede.

entre as cinco classes do TripAdvisor: 20478 horríveis (2,8%), 31522 ruins (4,3%), 121235 razoáveis (16,6%), 293823 muito boas (40,2%) e 263011 excelentes (36%); e o *Corpus* 7: 102390 avaliações divididas em 20478 para cada uma das 5 classes do TripAdvisor). Todos os sub-corpus, exceto os *Corpora* 1 e 6, tiveram suas avaliações embaralhadas aleatoriamente para evitar que os dados fossem concentrados, e as divisões de treinamento e teste poderiam capturar exemplos de todas as classes envolvidas. Foram feitos vários testes para determinar o número de épocas de treinamento para o modelo, e identificou-se que 30 épocas produzem bons resultados, e aumentar o número de épocas não resultou em maior precisão, mas aumento no tempo de processamento. Como forma de medição dos resultados, foram utilizadas as métricas: precisão e *recall*, e apresentando também a matriz de confusão.

Os resultados obtidos neste trabalho, superaram a maioria dos encontrados em Souza et al. (2018a) que foi usado de base para a construção dessa arquitetura. É interessante salientar que Souza et al. (2018a) utilizam uma base de dados menor tanto na quantidade de avaliações quanto do tamanho médio (68078 avaliações de tamanho máximo de 80 *tokens*). Além disso também foi feita a classificação para as cinco classes do Tripadvisor. A Tabela 4.1, apresenta os resultados obtidos para o *Corpus* 1, onde há um grande desbalanceamento das classes, 7,1% das avaliações

são negativas, 16,6% neutras e 76,2% positivas. Em relação a identificação das classes negativa e positiva obtivemos uma pequena melhora de 1,25% e 0,73%, mas para a classe neutra obtivemos uma melhora significativa de 8,91%. É interessante ressaltar que o desbalanceamento das classes em nosso trabalho e em Souza et al. (2018a) é praticamente o mesmo. Assim conseguimos melhorar significativamente a identificação da classe neutra que geralmente é mais difícil de ser identificada além de manter uma precisão aceitável para as outras duas classes.

Tabela 4.1: Resultados para o *Corpus* 1.

	Precisão e Recall		Matriz de confusão		
	Precisão	Recall	Negativa	Neutra	Positiva
Negativa	74,07%	66,65%	7021	3062	450
Neutra	60,07%	49,83%	2282	12057	9860
Positiva	91,2%	95,27%	240	5023	106019

A Tabela 4.2 apresenta os resultados obtidos com o *Corpus* 2, neste caso balanceamos a quantidade de avaliações positivas em relação às neutras, reduzindo-as a 121235. Os resultados obtidos indicaram melhora de 1,15% e de 15,05% de precisão para as classes negativa, e neutra em relação aos resultados encontrados para o *Corpus* 1, mas tivemos uma redução de 2,44% da precisão para a classe positiva. Em relação a Souza et al. (2018a) tivemos resultados próximos, melhores para as classes negativa e neutra (2,13% e 1,76% melhor respectivamente) e pior para a positiva (0,51% pior), sendo que a perda na classe positiva foi de menos de 1% e a melhora nas outras duas classes foi superior a isso, mantendo uma precisão global interessante.

Tabela 4.2: Resultados para o *Corpus* 2.

	Precisão e Recall		Matriz de confusão		
	Precisão	Recall	Negativa	Neutra	Positiva
Negativa	75,22%	68,94%	7243	3169	93
Neutra	75,12%	79,82%	2333	19369	2563
Positiva	88,76%	86,2%	81	3248	20792

A Tabela 4.3 retrata os resultados obtidos com o *Corpus* 3, onde balanceamos as três classes. Em relação aos resultados encontrados no *Corpus* 2, aqui obtivemos maior precisão para a classe negativa (aumento de 5,62%), porém as demais tiveram redução (3,13% e 0,28%). Considerando os resultados anteriores, ainda obtivemos boa precisão, porém em relação a Souza et al. (2018a) para classes balanceadas

obtivemos uma redução de 1,26% e de 3,89% para as classes negativa e positiva respectivamente e melhora de 4,56% para a classe neutra. Há uma melhora interessante na identificação da classe neutra utilizando a rede convolucional simplificada, porém perdemos esse ganho com a piora para as outras duas classes, talvez devido a redução de camadas que poderiam identificar melhor aspectos que identifiquem essas classes. Apesar disso, ainda consideramos os resultados relevantes pela simplificação da arquitetura e aumento da entrada.

Tabela 4.3: Resultados para o *Corpus* 3.

	Precisão e Recall		Matriz de confusão		
	Precisão	Recall	Negativa	Neutra	Positiva
Negativa	80,84%	83,21%	8677	1656	94
Neutra	71,99%	70,49%	1948	7292	1104
Positiva	88,48%	86,93%	139	1224	9066

A Tabela 4.4 apresenta os resultados obtidos com a as classes negativa e positiva balanceadas (excluindo as neutras) que se comparados a Souza et al. (2018a), houve pequena melhora de 0,7% e de 1,3% para as classes negativa e positiva respectivamente. Esses resultados apontam para uma boa capacidade de classificação da rede, observando a matriz de confusão nota-se a baixa taxa de erro e o valor de *recall* obtido fortalece o resultado.

Tabela 4.4: Resultados para o *Corpus* 4.

	Precisão e Recall		Matriz de confusão	
	Precisão	Recall	Negativa	Positiva
Negativa	96,76%	96,78%	10057	335
Positiva	96,79%	96,75%	338	10070

Na Tabela 4.5 foi utilizado o *corpus* novamente sem as avaliações neutras, porém mantendo o desbalanceamento original e novamente obteve-se resultados muito próximos e melhores aos encontrados em Souza et al. (2018a) melhora de 0,68% e de 0,46% para as classes negativa e positiva respectivamente, obtivemos melhora também no valor de *recall* para a classe negativa.

Diferente de Souza et al. (2018a), foram classificadas ainda as avaliações conforme as cinco classes do TripAdvisor. A Tabela 4.6 expõe os resultados encontrados utilizando a arquitetura desenvolvida. Nota-se que apenas para a classe 2 (ruim) obteve-se precisão inferior a 50%, sendo ineficaz para a identificação da classe, num geral não se obteve precisão alta (acima de 80%) para nenhuma das cinco classes.

Tabela 4.5: Resultados para o *Corpus* 5.

	Precisão e Recall		Matriz de confusão	
	Precisão	Recall	Negativa	Positiva
Negativa	89,62%	86,78%	9029	1376
Positiva	98,77%	99%	1111	110251

Para as classes 1 e 5 (horrível e excelente) a rede conseguiu classificar melhor, apesar de o *recall* para a classe 1 ser baixo. Observando a matriz de confusão, nota-se que ao tentar classificar estas duas classes a rede erra principalmente classificando as instâncias como classe 2 (para a classe 1) e como classe 4 (para a classe 5), algo que pode causar confusão até mesmo quando pessoas tentam realizar essa classificação como atestado posteriormente. Nota-se ainda que como o *corpus* é desbalanceado, o valor de *recall* para as classes mais abundantes de instâncias, classes 4 e 5, é maior.

Tabela 4.6: Resultados para o *Corpus* 6.

	Precisão e Recall		Matriz de confusão				
	Precisão	Recall	1	2	3	4	5
1	68,95%	38,48%	1582	1736	713	69	12
2	42,46%	38,1%	664	2447	3058	236	17
3	56,34%	58%	202	1584	14084	8002	326
4	61,59%	70,35%	22	126	5775	41197	11441
5	74,97%	65,68%	6	21	543	17521	34631

Por fim, utilizou-se uma partição do *corpus* onde todas as 5 classes estavam balanceadas, a Tabela 4.7 apresenta os resultados obtidos. Em relação ao observado na Tabela 4.6, ao balancear as classes obteve-se melhora significativa para as classes 1, 2 e 3 (2,03%, 7,79% e 7,62%, respectivamente), e piora para as classes 4 e 5 (8,27% e 4,57%, respectivamente). O *recall* obtido foi significativamente melhor para as classes 1, 2 e 3, e um valor inferior apenas para a classe 4. Devido o balanceamento das classes, as classes 4 e 5 possuíam menos representantes o que influencia na classificação, por isso a queda na precisão. Nota-se ainda que a maior dificuldade é identificar as classes 2 e 4 que não chegam a ser neutras e nem são os extremos de polaridade negativa e positiva, respectivamente, fenômeno este que pensou-se em ocorrer pela dificuldade natural de identificar essa diferença até mesmo por seres humanos, que inclusive pode ser observada na Tabela 4.8.

No intuito de demonstrar que de fato a determinação da polaridade de um texto não é uma tarefa trivial, foram recolhida uma amostra de avaliações presentes no *corpus* para que voluntários atribuíssem uma nota ente 1 e 5 (mesmo sistema de

Tabela 4.7: Resultados para o *Corpus* 7.

	Precisão e Recall		Matriz de confusão				
	Precisão	Recall	1	2	3	4	5
1	70,98%	64,31%	2636	1336	103	14	9
2	50,25%	61,52%	931	2499	565	56	11
3	63,96%	52,2%	132	1026	2175	748	86
4	53,32%	58,81%	14	89	488	2362	1063
5	70,4%	66,85%	6	27	80	1256	2763

notas que os viajantes atribuem no TripAdvisor). Foram coletadas arbitrariamente 385 avaliações, sendo 77 de cada uma das 5 classes do TripAdvisor (horrível, ruim, razoável, muito bom e excelente), o que nos dá um grau de confiança de 95% e margem de erro de 5%, por entender que seriam valores representativos, conforme a definição dessas métricas apresentadas em Bessegato (2012). Para esta tarefa, 5 avaliadores voluntários atribuíram a nota que acharam condizente com o texto escrito pelo viajante. Essa amostragem também foi realizada como um comparativo em relação aos resultados obtidos com a arquitetura de rede neural profunda desenvolvida.

A Tabela 4.8 apresenta a precisão, o *recall* e a matriz de confusão gerada a partir das notas atribuídas pelos voluntários para os comentários selecionados, as linhas numeradas de 1 a 5 correspondem às 5 classes mencionadas anteriormente. Apontando que mesmo pessoas têm dificuldade para determinar uma classe conforme o conteúdo do comentário, visto que, opinião é algo subjetivo. Nota-se que quando o comentário está entre um dos extremos (horrível ou excelente) as pessoas têm mais facilidade de identificá-los porém entre as outras três classes intermediárias a dificuldade é maior. Na Tabela 4.7 apresentamos o resultados obtidos pela arquitetura de rede convolucional desenvolvida com um conjunto de avaliações balanceadas entre as 5 classes do TripAdvisor e observa-se que o método proposto possui precisão e acurácia melhor para todas as 5 classes. É relevante ressaltar que a diferença é bastante significativa: 13,94%, 11,32%, 17,73%, 16,56% e 11,3% para as classes 1, 2, 3, 4 e 5, respectivamente. Essas evidências apontam que o método é capaz de classificar avaliações de hotéis com precisão relativamente aceitável, vista a classificação em 5 classes.

Como nesse trabalho foi feita a mesclagem das classes positivas e negativas entre si, o mesmo foi feito com a classificação realizada pelos voluntários, deixando assim três classes: negativa, neutra e positiva. A Tabela 4.9 apresenta o resultado dessa mesclagem, o que fortalece a ideia de que quanto menor o número de classes é

Tabela 4.8: Classificação da amostra por voluntários conforme as cinco classes possíveis do TripAdvisor.

	Precisão e Recall		Matriz de confusão				
	Precisão	Recall	1	2	3	4	5
1	57,04%	40%	154	148	42	33	8
2	38,93%	37,92%	99	146	80	53	7
3	46,23%	47,79%	9	66	184	96	30
4	36,76%	50,13%	6	12	73	193	101
5	59,1%	54,8%	2	3	19	150	211

mais fácil para que uma pessoa defina a qual classe o comentário pertence. Uma vez que um dado comentário pode ser considerado "muito bom" para um indivíduo e "excelente" por outro e vice-versa, mas a diferença entre negativo e positivo é mais fácil de ser identificado. Ao comparar a classificação pelas pessoas com a realizada pela rede com as 3 classes balanceadas observa-se que as pessoas obtiveram maior precisão para classificar as avaliações negativas (3,97%), porém com *recall* 11,41% menor ao alcançado pela rede. Em relação às classes neutra e positiva, o método proposto obteve precisão 25,76% e 14,22% maior, respectivamente. Observa-se ainda a grande dificuldade que as pessoas tiveram para classificarem as avaliações que possuíam polaridade neutra, sendo que o método desenvolvido atingiu um valor razoável devida a dificuldade de identificá-la, conforme a literatura.

Tabela 4.9: Classificação da amostra por voluntários mescladas em 3 classes.

	Precisão e Recall		Matriz de confusão		
	Precisão	Recall	Negativa	Neutra	Positiva
Negativa	84,81%	71,8%	547	122	101
Neutra	46,23%	47,79%	75	184	126
Positiva	74,26%	85,06%	23	92	665

É interessante ressaltar que como a rotulagem das *reviews* foram feitas pelos próprios viajantes, o que também é subjetivo. A Figura 4.2 apresenta uma das avaliações selecionadas arbitrariamente para a classificação pelos voluntários onde o usuário deu uma nota 1 (horrível) para o hotel, porém quando alguém lê a avaliação geralmente não tende a dar uma nota negativa. Casos como este ocorrem no *corpus*, o que causa confusão tanto para pessoas quanto para algoritmos de aprendizagem pela contradição encontrada, se tornando mais uma adversidade com a qual o modelo proposto precisou lidar.

```
Foi a segunda vez que nos hospedamos nesse local. Fizeram
algumas mudanças como oferecem jantar, máquina de café
expresso no café da manhã, tem a opção de meia pensão. Os
funcionários educados e prestativos. O valor da diária
muito bom e uma qualidade nos serviços.
```

Figura 4.2: Avaliação com nota 1 atribuída pelo viajante, porém não possui conotação negativa para a maioria das pessoas.

4.6 Conclusão

Este trabalho apresentou o desenvolvimento de uma arquitetura de rede neural convolucional que é capaz de classificar avaliações de hotéis escritas em língua portuguesa, mapeando cada avaliação para duas, três e cinco classes distintas com precisão atrativa. A abordagem é atrativa devido aos resultados obtidos e à arquitetura não muito complexa da rede convolucional para classificar avaliações de hotéis no nível de documento.

Comparamos os resultados obtidos neste trabalho aos de Souza et al. (2018a) que realizaram experimento semelhante para uma base de dados com avaliações com tamanhos significativamente menores (maior avaliação com 80 *tokens*) e com número total de avaliações muito menor (68078 avaliações), porém com um desbalanceamento entre as classes parecido, visto que ambas as bases foram recolhidos do site TripAdvisor. Obtivemos resultados um pouco melhores em praticamente todas as subdivisões que fizemos do *corpus*, utilizando uma arquitetura baseada no trabalho de Souza et al. (2018a), mas com uma rede menos complexa e mesmo assim obtivemos bons resultados. É importante ressaltar que utilizamos como camada de saída um número de neurônios de acordo com a quantidade de classes que classificaríamos utilizando *softmax* como função de ativação e em Souza et al. (2018a) foi utilizado apenas um neurônio na camada de saída para classificar em 2 ou 3 classes utilizando a função de ativação *sigmoid* e posteriormente dividindo as faixas conforme as classes.

Comparamos ainda os resultados obtidos pela nossa rede aos encontrados pela classificação de avaliações realizada por voluntários, o que mostrou que nosso método obteve melhor precisão para classificar as avaliações em 5 classes para todas as 5 classes. Quando consideramos as três classes (negativa, neutra e positiva) balanceadas, nossa rede obteve precisão consideravelmente melhor para as classes neutra e positiva (25,76% e 14,22%, respectivamente) e um pouco pior para a classe negativa (3,97%) porém com *recall* menor, cerca de 11%. Resultado que aponta que nosso modelo possui a capacidade de classificar, com precisão relevante, avaliações

de hotéis escritas em língua portuguesa.

Como trabalhos futuros, seria interessante observar o efeito das classes gramaticais na classificação das avaliações e identificar quais são as classes gramaticais que mais contribuem para a determinação de polaridade além dos adjetivos que naturalmente têm esse papel. Outra possibilidade seria a construção de uma rede neural recorrente, LSTM por exemplo, evitando a criação de matrizes que podem possuir muitos zeros causando grande consumo de memória.

Capítulo 5

Conclusões Gerais e Trabalhos Futuros

Este trabalho teve como objetivo o desenvolvimento de modelo que utilize redes neurais artificiais, mais especificamente modelos baseados nas arquiteturas relacionadas com o aprendizado profundo, para a tarefa de classificação de polaridade de avaliações de hotéis escritas em Língua Portuguesa. Para alcançar esse objetivo foram desenvolvidos dois modelos de redes neurais convolucionais para realizar a tarefa de classificação de polaridades.

O primeiro modelo consistiu em um rede convolucional que possuía 15 camadas: 1 camada de *embedding* (entrada dos dados), 4 camadas convolucionais, 4 camadas de *pooling*, 3 camadas de *dropout* e 3 camadas densas, tendo a camada de saída com 1 neurônio utilizando *sigmoid* como função de ativação. Esse modelo obteve resultados relevantes alcançando precisão acima de 95% para as classificação das classes positiva e negativa isoladamente e mantendo o balanceamento entre as classes, e acima de 90%, 67% e 80% para as classes positiva, neutra e negativa, respectivamente. Os resultados obtidos nesse trabalho foram comparados aos de Martins et al. (2017) apontando o modelo como boa solução por apresentar resultados superiores em diversos casos, além da capacidade de identificar com precisão considerável a classe neutra, que é de difícil identificação conforme a literatura.

O segundo modelo construído foi baseado no primeiro, porém com adequações de hiper-parâmetros, redução e realocação de camadas, utilização de um vetor representação de palavras maior (200 dimensões) e de um *corpus* cerca de 10 vezes maior em quantidade de *reviews* e com número de *tokens* também superior. A arquitetura desenvolvida possuía 10 camadas: 1 camada de *embedding*, 2 camadas convolucionais, 2 camadas de *pooling*, 2 camadas de *dropout* e 3 camadas densas, tendo a

camada de saída variável conforme a quantidade de classes que seriam classificadas podendo conter 2, 3 ou 5 neurônios (para classificação em negativo e positivo; negativo, neutro e positivo; e as 5 classes do TripAdvisor, respectivamente), utilizando *softmax* como função de ativação. Essas modificações na estrutura permitiram obter resultados ainda melhores do que os observados no primeiro modelo, salvo em 3 ocasiões dentre 13 possíveis. Diferente do primeiro modelo, neste segundo houve a capacidade de classificar as avaliações conforme as 5 classes e a partir de amostra recolhida do *corpus* classificada por voluntários, um comparativo foi realizado entre os resultados alcançados utilizando o modelo proposto e a classificação gerada pelas pessoas, apontando que apenas em um caso as pessoas obtiveram maior precisão do que o modelo desenvolvido. Resultado que aponta que o modelo possui a capacidade de classificar, com precisão relevante, avaliações de hotéis escritas em língua portuguesa.

Como era preciso uma base de dados maior e com a mesma estrutura da que utilizada para testar o primeiro modelo, desenvolveu-se um *corpus* que se tornou um sub-produto dessa pesquisa, que está disponível para a comunidade.[1]

O desenvolvimento desses modelos contribuem para o avanço do estado da arte no que diz respeito à utilização de redes neurais artificiais, mais especificamente redes neurais convolucionais para a classificação de polaridade de *reviews* de hotéis escritas em Língua Portuguesa. Além disso podem ser utilizados como base para o desenvolvimento de ferramentas capazes de ajudar pessoas a retirar informação subjetiva de texto da *Web*.

5.1 Trabalhos Futuros

A identificação da classe neutra continua um desafio pela dificuldade de ambos os modelos em identificá-la, apesar de apresentarem resultados relativamente bons, por isso novas arquiteturas podem ser construídas no intuito de aumentar a capacidade de identificação da classe. Além da classificação em 5 classes que obteve resultados melhores em relação à realizada pelos voluntários, mas que pode ser melhorada, visto que em nenhuma classe foi alcançada precisão superior a 74,97%. Outro ponto a ser analisado seria a realização de testes com outras línguas, como o inglês e o espanhol por exemplo, para avaliar a precisão do método para idiomas diferentes do português. Na linha de testar o método em outras línguas, um outro passo seria realizar um comparativo utilizando bases de dados amplamente conhecidas

[1] Os arquivos do *corpus* estão disponíveis em: https://bit.ly/2JVRJbI

na literatura, como a *Sentiment Treebank*, para averiguar a precisão do modelo em relação a outros trabalhos utilizando a mesma base.

A utilização de redes neurais recorrentes é sugerida quando não se tem um tamanho fixo da entrada, como o caso das avaliações, neste trabalho foi utilizado o preenchimento com zeros para solucionar este problema, o que tende a resultar em maior tempo de processamento. Com isso o desenvolvimento de um modelo que utilize este tipo de rede possa alcançar bons resultados reduzindo este tempo.

Referências Bibliográficas

Aciar, S. (2010). Mining context information from consumers reviews. Em *Proceedings of Workshop on Context-Aware Recommender System. ACM*, volume 201.

Agarwal, B.; Mittal, N. et al. (2016). *Prominent feature extraction for sentiment analysis*. Springer.

Ain, Q. T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B. & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, 8(6):424.

April, Y. & Daryl, C. (2015). Multiclass sentiment prediction using yelp business reviews. Em *CS224d: Deep Learning for Natural Language Processing, Reports*.

Attia, M.; Samih, Y.; Elkahky, A. & Kallmeyer, L. (2018). Multilingual Multiclass Sentiment Classification Using Convolutional Neural Networks. Em chair), N. C. C.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S. & Tokunaga, T., editores, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Baccianella, S.; Esuli, A. & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Em *Lrec*, volume 10, pp. 2200--2204.

Becker, K. & Tumitan, D. (2013). Introdução à mineração de opiniões: Conceitos, aplicações e desafios. *Simpósio Brasileiro de Banco de Dados*.

Bessegato, L. F. (2012). Estimativas e tamanhos de amostras. http://www.bessegato.com.br/PUC/iec_transp_09.pdf. Acesso em: 12 de jul. 2018.

Bildhauer, F. & Schäfer, R. (2013). Token-level noise in large web corpora and non-destructive normalization for linguistic applications. *Proceedings of Corpus Analysis with Noise in the Signal (CANS 2013)*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H. & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`. Accessed: 2018-01-30.

Cieliebak, M.; Deriu, J. M.; Egger, D. & Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis. Em *5th International Workshop on Natural Language Processing for Social Media, Boston, MA, USA, December 11, 2017*, pp. 45--51. Association for Computational Linguistics.

de Oliveira Gonçalves, P. (2015). Um benchmark para comparação de métodos para análise de sentimentos.

Dinu, L. P. & Iuga, I. (2012). The naive bayes classifier in opinion mining: in search of the best feature set. Em *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 556--567. Springer.

dos Santos, A. G. L.; Becker, K. & Moreira, V. (2014). Um estudo de caso de mineração de emoções em textos multilíngues. *BRASNAM - III Brazilian Workshop on Social Network Analysis and Mining*.

dos Santos, C. & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. Em *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69--78.

Duran, M. S.; Avanço, L.; Aluísio, S.; Pardo, T. & Nunes, M. d. G. V. (2014). Some issues on the normalization of a corpus of products reviews in portuguese. Em *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 22--28.

Erik Cambria, A. H. (2015). *Sentic Computing*, volume volume 1 of *Socio-Affective Computing 1*. Springer, Springer Cham Heidelberg New York Dordrecht London, 1 edição.

Esuli, A. & Sebastiani, F. (2007). Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, 17:1--26.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Guzman, E. & Maalej, W. (2014). How do users like this feature? a fine grained sentiment analysis of app reviews. Em *Requirements Engineering Conference (RE), 2014 IEEE 22nd International*, pp. 153--162. IEEE.

Hartmann, N. S.; Avanço, L. V.; Balage Filho, P. P.; Duran, M. S.; Nunes, M. D. G. V.; Pardo, T. A. S.; Aluisio, S. M. et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. Em *International Conference on Language Resources and Evaluation, 9th*. European Language Resources Association-ELRA.

Hassan, A. & Mahmood, A. (2017). Deep learning approach for sentiment analysis of short texts. Em *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*, pp. 705–710.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735--1780.

Jesus, A. (2014). História das redes sociais: do tímido classmates até o boom do facebook. http://www.techtudo.com.br/artigos/noticia/2012/07/historia-das-redes-sociais.html. Acesso em: 12 de jul. 2018.

Jurafsky, D. & Martin, J. H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Jurafsky, D. & Martin, J. H. (2017). Speech and language processing.

Kasper, W. & Vela, M. (2011). Sentiment analysis for hotel reviews. Em *Computational linguistics-applications conference*, volume 231527, pp. 45--52.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333--347.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627--666.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1--167.

Martins, G. S.; Oliveira, A. d. P. & Moreira, A. (2017). Sentiment analysis applied to hotels evaluation. Em *International Conference on Computational Science and Its Applications*, pp. 710--716. Springer.

Meyer, C.; Grabowski, R.; Han, H.-Y.; Mantzouranis, K. & Moses, S. (2003). The world wide web as linguistic corpus. *Language and Computers*, 46:241--254.

Mikolov, T.; Chen, K.; Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OTHERO, G. d. Á. & MENUZZI, S. d. M. (2005). Linguística computacional: teoria e prática. *São Paulo: Parábola*.

Pak, A. & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Em *LREc*, volume 10, pp. 1320--1326.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1--135.

Patil, P. (2017). Application for data mining and web data mining challenges. *International Journal of Computer Science and Mobile Computing*, 6(3):39--44.

Pennington, J.; Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. Em *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532--1543.

Petrović, S.; Osborne, M. & Lavrenko, V. (2010). The edinburgh twitter corpus. Em *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pp. 25--26.

Rocha, P. A. & Santos, D. (2000). Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*.

Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V. & Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. Em *Proceedings of LREC*.

Sharath, T. & Tandon, S. (2017). Topic based sentiment analysis using deep learning. *arXiv preprint arXiv:1710.10498*.

Shimada, K.; Inoue, S.; Maeda, H. & Endo, T. (2011). Analyzing tourism information on twitter for a local city. Em *Software and Network Engineering (SSNE), 2011 First ACIS International Symposium on*, pp. 61--66. IEEE.

Shirani-Mehr, H. (2014). Applications of deep learning to sentiment analysis of movie reviews. Em *Technical Report*. Stanford University.

Sodanil, M. (2016). Multi-language sentiment analysis for hotel reviews. Em *MATEC Web of Conferences*, volume 75, p. 03002. EDP Sciences.

Souza, J. G. R. d.; de Paiva Oliveira, A.; de Andrade, G. C. & Moreira, A. (2018a). A deep learning approach for sentiment analysis applied to hotel's reviews. Em *International Conference on Applications of Natural Language to Information Systems*, pp. 48--56. Springer.

Souza, J. G. R. d.; de Paiva Oliveira, A. & Moreira, A. (2018b). Development of a brazilian portuguese hotel's reviews corpus. Em *International Conference on the Computational Processing of Portuguese*. Springer.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929--1958.

Tang, D.; Wei, F.; Qin, B.; Liu, T. & Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification. Em *SemEval@ COLING*, pp. 208--212.

Tsytsarau, M. & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478--514.

Valdivia, A.; Luzón, M. V. & Herrera, F. (2017a). Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews? Em *International Conference on Hybrid Artificial Intelligence Systems*, pp. 15--25. Springer.

Valdivia, A.; Luzón, M. V. & Herrera, F. (2017b). Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews? Em *International Conference on Hybrid Artificial Intelligence Systems*, pp. 15--25. Springer.

Vateekul, P. & Koomsubha, T. (2016). A study of sentiment analysis using deep learning techniques on thai twitter data. Em *Computer Science and Software*

Engineering (JCSSE), 2016 13th International Joint Conference on, pp. 1--6. IEEE.

Yogatama, D.; Dyer, C.; Ling, W. & Blunsom, P. (2017). Generative and discriminative text classification with recurrent neural networks. *arXiv preprint arXiv:1703.01898*.

Zhang, L.; Wang, S. & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1253.

Zhang, Y. & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.