Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Um Estudo Comparativo entre Abordagens baseadas em Sistemas *Fuzzy* e Redes Neurais Artificiais para Estimar a Importância de Comentários sobre Produtos e Serviços

Roney Lira de Sales Santos

Teresina-PI, 21 de Março de 2017

Roney Lira de Sales Santos

Um Estudo Comparativo entre Abordagens baseadas em Sistemas *Fuzzy* e Redes Neurais Artificiais para Estimar a Importância de Comentários sobre Produtos e Serviços

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para aprovação na disciplina Dissertação de Mestrado.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Raimundo Santos Moura

Teresina-PI

21 de Março de 2017

S237e Santos, Roney Lira de.
Um Estudo Comparativo entre Abordagens baseadas em Sistemas *Fuzzy* e Redes Neurais Artificiais para Estimar a Importância de Comentários sobre Produtos e Serviços/ Roney Lira de Sales Santos. – Teresina-PI, 2017.
125 f. : il.: color.

Dissertação (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, 2017.
Orientador: Prof. Dr. Raimundo Santos Moura

1. Mineração de Opinião. 2. Processamento de Linguagem Natural. 3. Redes Neurais Artificiais. 4. Sistemas *Fuzzy*. I. Raimundo Santos Moura. II. Universidade Federal do Piauí. III. Um Estudo Comparativo entre Abordagens baseadas em Sistemas *Fuzzy* e Redes Neurais Artificiais para Estimar a Importância de Comentários *Web*.

CDD 004.21

Um Estudo Comparativo entre Abordagens baseadas em Sistemas Fuzzy e Redes Neurais Artificiais para Estimar a Importância de Comentários sobre Produtos e Serviços

RONEY LIRA DE SALES SANTOS

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Centro de Ciências da Natureza da Universidade Federal do Piauí, como parte integrante dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Aprovada por:

Prof. Raimundo Santos Moura

(Presidente da Banca Examinadora)

Prof. Thiago Alexandre Salgueiro Pardo

(Examinador Externo)

Prof. Pedro de Alcântara dos Santos Neto

(Examinador Interno)

Prof. Ricardo de Andrade Lira Rabêlo

(Examinador Interno)

Teresina, 21 de março de 2017.

Aos meus pais Maria de Fátima e Raimundo Nonato, ao meu irmão Anderson Lira, aos meus avós, tios, primos, amigos e professores, por fazerem parte de cada dia que eu trabalhei para conquistar esse mérito.

Agradecimentos

Em um dos dias que mais me emocionei na minha vida, alguns amigos da minha inesquecível turma que tive o prazer de fazer parte no ensino médio me disse: “o Roney prefere receber um obrigado a dizer um”. Claro que o sentido foi no qual eu sempre me dispus a ajudar qualquer um que me pedisse, mesmo eu sempre pedindo ajuda também a eles nas tarefas diárias. Porém, chegou um outro momento: onde não vou economizar nos agradecimentos. Eu tenho tantas pessoas pra agradecer que eu confesso que demorei para fazer esse texto, pois gostaria que ficasse documentado o meu sentimento verdadeiro quanto a elas, assim como o meu relacionamento com todas essas pessoas.

Primeiramente gostaria de agradecer a minha mãe, D. Maria de Fátima, a Fafá. É imensurável o meu amor, meu respeito, minha admiração por esta pessoa. Todos os ensinamentos, toda a vivência, toda a proteção que ainda hoje tem para comigo é algo que eu sempre levei e sempre vou levar na minha vida. Nenhuma frase que eu escreva aqui vai mostrar o meu agradecimento suficiente que ela mereça. Não menos importante, agradeço ao meu pai, Sr. Raimundo Nonato por manter firme seus pensamentos e crenças. Saiba que mesmo que alguns dos nossos pensamentos sejam diferentes (é natural), cresci e sou o homem íntegro e responsável que sou por causa de sua criação correta e firme. Ao meu irmão Anderson Lira (Sales, como ele quer que seja chamado) por todo o companheirismo que sempre tivemos e o respeito mútuo entre nós, fruto da criação fantástica de nosso pai e nossa mãe. Meu futuro Mestre!

Palavras adequadas faltam para agradecer aos meus avós maternos, D. Alexandrina Rodrigues e Sr. Raimundo Nonato Sales. Quando estava prestes a realizar meu sonho de cursar o curso que sempre quis, eles abriram as portas da sua casa para me acolher e como sempre digo: lá é a minha segunda casa. Credito a minha criação tanto na infância quanto na fase adulta a eles. Aos meus avós paternos, D. Maria Lira e Sr. Aldenor dos Santos (*in memoriam*), que por mais que não tenha passado tanto tempo juntos, são essenciais na minha vida, tanto fisicamente quanto na minha memória.

Como não ser feliz quando se tem uma família unida? Eu dou graças ao meu bom Deus sempre por ter, além dos meus pais, meu irmão, meus avós já citados, os meus tios e tias, primos e primas. Tenho o prazer de ser da mesma família que vocês! Tia Corrinha (Socorro Sales), minha primeira professora; Tia Rosângela, que eu serei o bebê dela para sempre; Tia Evinha (Eva Sales), Tia Nilsa (Francisca Sales), Tia Zélia, Tio Adão, Tio Josias, Tio Antônio, Tio Messias, Tia Noêmia e Tio Lirinha (Pedro Lira), meu companheiro de viagens. Orgulho de ter o mesmo sangue que vocês! Sem esquecer dos meus tios e tias frutos dos casamentos: Tio Antônio (sim, outro hahaha), Tio Freitas, Tia Jocélia (Ié,

minha primeira babá), Tia Patrícia, Tia Noninha (Francilande Araújo) e a minha mais nova tia, Tia Andréia! Obrigado por tudo!

Meus primos... aah meus primos! Meus irmãos de mães diferentes! Nathália Sales, a prima da mesma época, quase matamos a vovó Alexandrina do coração quando fomos atrás de algodão-doce (?) quando tinhamos 2 anos; Igor Sales, talvez meu primo mais próximo; Nayana Sales, sem palavras pra ti, minha prima mais descolada; Iasmim Sales, minha priminha mais carinhosa comigo; Giovanna Sales, minha prima de gênio forte (como eu) e determinada; Lucas Sales e Luanne Sales, meus priminhos mais novos; Minhas primas-irmãs Juliana e Joyce Carvalho, por todo o companheirismo desde nossa infância e ainda vivemos uns do lado do outro, né; João Paulo, santista que nem a família, meu companheiro de viagem também; Minha priminha Paula Joseane, Janayna Lira, mesmo de longe, fez parte de uma fase importante da minha formação. Aos meus primos que não são de sangue, mas são de coração, né Edinaldo Cardoso, Guilherme Rodrigo, Rafael Santiago, Samuelson Sales, Mayara Jéssica, Mayrla Dias e Danilo Brito, "tamo junto"! Claro que não posso deixar de mencionar a minha cunhada Eduily Vaz, que já faz parte na nossa família, adorada por mim e por todos! Nossa, não tenho como agradecer vocês por tudo!

Tenho a família de sangue, mas também se criou uma família de trabalho, de estudo, de companheirismo, enfim... de amizade! Eu agradeço a Deus todos os dias por ter amigos de verdade. Meus amigos do ensino fundamental, ensino médio, ensino superior, agradeço a vocês por estarem sempre aturando a minha chatice! Alguns eu não podia deixar de mencionar neste capítulo de agradecimentos, não seria justo não fazer isso, por isso, abre parágrafo!

Agradeço demais aos meus amigos da turma 2011.1 por todo o companheirismo que perdura nesses 6 anos. Alguns mais distantes, outros mais perto, mas lembro com carinho de todos! Obrigado Alan Santos, o empreendedor; Allan Melão, o resolve-tudo, companheiro de laboratório, mito demais; Ângelo Siqueira, Carlos Henrique, Daniel Vieira, Francisco Assis, o mito; Jales Roberto, fanático por vôlei; Jefferson Chaves, companheiro de laboratório; Joaquim Alves, o cara que eu discutia HIMYM[1] toda semana; Josimar Alves, Laércio Andrade, o cara da tecnologia; Lucas Sena, meu companheiro de PES[2] e de laboratório; Paulo Sérgio, o cara que me recomenda só músicas 10 de 10; Rafael Moreira, o dono da casa do pingue-pongue; Rafael Maciel, o dono da casa dos churrascos (tá, foi só um, mas já tá definido né haha); Ramiro Cavalcanti, Rayllson Nascimento, o rico da turma; Samuel Feitosa, o cara mais sensacional da turma e Selles Gustavo, o cara que eu mais me espelho, sou fã demais. Essa turma vai ficar marcada na minha história pra sempre!

Claro que numa turma há um relacionamento mais próximo com algumas pessoas.

---

1 *How I Met Your Mother*, série de TV
2 *Pro Evolution Soccer*, jogo virtual

Amizade de vai durar muito, mas muito tempo! Agradeço aos meus companheiros de Mestrado Francisco Bruno (o Bill!) e Ítalo Brasileiro por estarmos sempre em contato e sempre um resolvendo o problema do outro. É o fruto que a 2011.1 gerou. Vocês vão ter muito sucesso, eu acredito! Ao meu calouro de Mestrado e companheiro de laboratório João Paulo Albuquerque, um cara sensato, de opinião forte e firme, que merece todo o sucesso do mundo e que me ajudou pra caramba nesse tempo de convivência no LPLN[3], muito obrigado. Ao meu ex-companheiro de laboratório Jardeson Leandro, onde sofremos juntos uma boa parte dos trabalhos da graduação e da Iniciação Científica e temos muitos projetos ainda pra fazer, vou cobrar! Por fim, mas não menos importante, eles sabem disso, o meu muito obrigado ao Saulo de Társio e ao Ruhan Bello, por todos os momentos de descontração (e sofrimento, precisa nem dizer né hahaha) que passamos, junto com o Ítalo Brasileiro. Nossa amizade é forte e eu agradeço demais a vocês!

Felizmente eu conquistei amigos também fora da minha turma. São pessoas que me ajudaram e me ajudam diariamente, se não com o meu trabalho em si, mas com o meu bom humor para com ele. Agradeço aos meus ex-companheiros de EASII[4] Ronyérison Braga, Gleison Andrade, Rafael Fontinele, Wermeson Reis, Thasciano Carvalho e Werney Lira, além dos meus amigos de outros laboratórios Wender Reis, Luis Guilherme, Francisco Neto, Ramon Nepomuceno, Laysson Oliveira, Jonathas Evangelista, Francisco Júnior (Juninho) Ranulfo Plutarco e Moisés Bispo. Também não esqueço de agradecer e desejar boa sorte aos meus companheiros de Mestrado da turma de 2015 Alan Santos, Denise Costa, Dennis Sávio, Francisco Imperes, Hugo Cordeiro, Ivenilton Moura, Jaílson Leocádio, Jurandir Cavalcante, Marcos Frazão, Martony Demes e Sebastião Galeno por todo o tempo que passamos nas disciplinas e comentando sobre nosso futuro. Sem faltar com a consideração e respeito que criei pelo Carlos Sá, meu companheiro de laboratório, de pesquisa e de orientador! Aprendi muita coisa com você! Além disso, ainda tem os mestres Pedro Almir, Jefferson Henrique e Mateus Campanhã, os quais sou fã e que me espelho fortemente por todo o legado que deixaram no PPGCC, além do meu mestre amigo Rogério Figueredo de Sousa que foi de suma importância para a realização do trabalho com sua ideia inovadora que é a base deste trabalho. Ao meu aluno nota 10 Raul Pereira, que fez parte de um dos meus melhores momentos nesses dois anos: o prazer de lecionar. Obrigado, pessoal!

Além de todas essas pessoas que tenho o prazer de ser amigo, durante esse tempo eu fui abençoado com pessoas que vivem o meu dia-a-dia, aqueles amigos mais próximos. Obrigado Jordão Frazão e Giovanni Emanuel por me aguentarem e por fazer parte de provavelmente as melhores lembranças que tenho do meu tempo de UFPI, junto com o Saulo de Társio (vários nomes pra esse grupo foram dados, então prefiro não definir um hahaha). Muito obrigado Clenildo Luz, Isaías Miranda, Onofre Neto, Stéfano Carvalho

[3] Laboratório de Processamento de Linguagem Natural, nome extra-oficial adotado pelos integrantes do Laboratório de Computação Científica

[4] Laboratório de Engenharia de Software e Informática Industrial

e Vanderson Moura, meus irmãos do *Brotherhood*, que fizeram parte ativamente desse meu tempo no Mestrado, sempre a parte do projeto e sempre ajudando com o carinho e amizade, muito obrigado mesmo. Aos meus grandes amigos do ensino médio Sávio Caiubi, Wilson Dasein e João Guilherme (companheiro na turma 2011.1) por serem presentes durante esse tempo, muito obrigado!

Três caras merecem um destaque maior. Pelo o que representaram e que representam durante esse caminho todo. Posso dividir esses dois anos de Mestrado com uma parte importante para cada um deles. Primeiro, lá no início, devo ao Hugo Portela todo o companheirismo, amizade, paciência (muita, hein) e ajuda em tudo. Um cara que você pode contar pra qualquer coisa que ele vai estar lá lhe ajudando. Agradeço demais tua amizade comigo, cara. Alí na definição do projeto e começo do desenvolvimento, devo ao Joselito Júnior a honra de me aguentar ao questionar tudo que dava certo e tudo que não dava certo. Nossas conversas eram (e são até hoje) sempre alegres e de alto nível (hahahaha). Cara, te agradeço demais por ser meu amigo e por me ajudar nesse processo. Por fim, qualificação e defesa, devo muito ao Gilvan Veras, meu irmão de todas as horas mesmo, nem que seja 1h da manhã estávamos lá. Meu companheiro de laboratório, de saídas, de problemas e sucessos e que fiquei feliz ao ver o sucesso dele. Devo muito, mas muito a ele essa conquista. Não são três amigos, são três irmãos que eu ganhei pra vida toda. Além disso, eles ainda me deram de presente três cunhadas: Larissa Moura, Lara Susan e Larissa Viana, respectivamente. Mereço isso tudo? Muito, mas muito obrigado mesmo.

Tive o prazer de estudar e trabalhar numa instituição íntegra e correta. Agradeço muito à Universidade Federal do Piauí pelo espaço que me fez crescer pessoalmente, academicamente e profissionalmente. Muito obrigado ao Sr. Robert Reis, Anathália Cristina, Delson Bonfim, Constança Dolores e a todo o pessoal do NTI-UFPI[5], pois aprendi muito no tempo que convivi diariamente nesse espaço, conhecimento esse que levei por todo esse período na universidade. Sou grato também pelo tempo que passei com o pessoal do NUPLID[6], ao Prof. Saulo Brandão, Shisleny Lopes, Priscila Viviane, Ademar Júnior, Alessandro Márcio, Lucas Rolim, por todo os ensinamentos mútuos que realizamos neste período. Não menos importante, gostaria de agradecer a cada Professor que fez parte da minha formação na graduação e na pós-graduação, pois sem eles eu não seria o que sou hoje: Prof. Paulo Sérgio, Prof.ª Vânia Barbosa, Prof.ª Maria Inês, Prof. André Macêdo, Prof. Rodrigo Veras, Prof. Érico Leão, Prof. Jurandir Lopes, Prof. Ivan Saraiva, Prof. Alexandre Nojoza, Prof. Antônio Macêdo, Prof. Kelson Aires, Prof. Pessoa Júnior, Prof. Francisco Nilson, Prof.ª Jackelya Araújo, Prof. Pedro Alcântara, Prof. Luiz Cláudio Demes, Prof. Magno Alves, Prof. Gildásio Guedes, Prof. Antônio Costa, Prof. Flávio Ferry, Prof. Weslley Emmanuel, Prof.ª Rosianni Cruz, Prof. Vinícius Machado, Prof. Francisco Vieira

[5] Núcleo de Tecnologia da Informação
[6] Núcleo de Pesquisa em Literatura Digitalizada

e Prof. André Castelo Branco.

Gostaria de dar um destaque para três dos meus mestres nesta caminhada: Prof. Armando Soares, meu tutor nas disciplinas de Engenharia de Software I e II e Estágio Supervisionado, onde aprendi não só os assuntos ministrados nessas disciplinas. Seus ensinamentos me fizeram ser um profissional e um acadêmico muito melhor do que eu era antes, além disso, de ser uma pessoa certa e pontual para com as responsabilidades acadêmicas e por que não pessoal; e ao Prof. Ricardo Lira, meu tutor nas disciplina de Introdução à Redes Neurais Artificiais, na qual foi uma das minhas preferidas durante a graduação e por isso dá nome a meu trabalho de Mestrado e meu co-orientador extra-oficial, pois além de ser uma pessoa focada e comprometida com o belo trabalho que realiza com seus alunos, é uma pessoa de caráter fantástico, na qual também me espelho muito. Muito obrigado!

Já um dos três mestres merece um parágrafo inteiro. Professor Doutor Raimundo Santos Moura, muito obrigado por tudo. Por todos esses 4 anos de orientação, de dedicação, de confiança, de paciência, de ensinamentos, de ideias, de artigos aceitos, de comemorações, enfim, de sucesso. Nosso contato irá perdurar por muito mais tempo, pois não tenho o direito de não ter a opinião do meu orientador para meus trabalhos futuros. Muito obrigado mesmo por todo esse tempo e por esta Dissertação, que não é só minha, é nossa. Sou grato por tudo, obrigado.

Como um amante de esportes, sempre estive acompanhando vários deles durante o (pouco) tempo livre neste período. Nisso, também fiz vários amigos que, diretamente e indiretamente, me ajudaram no meu trabalho, mesmo de longe. Ao meus amigos de São Caetano do Sul, torcedores do AD São Caetano, meu time do coração, da velha e nova guarda: Felippe Freitas, Felipe Godoy, Raphael Cassiolato, Fábio Aleixo, Fábio Augusto, Luiz Felipe, Ronie James, Denys Lima, André Santana, Priscila Iurtchechen, Ingrid Oliveira, Diego Muñoz, Fernando Paes, Alan Andrade, Leandro Clobochar, Renan Sant'Ana, Carlos Lazarini, Wedson Leal, Timo Domenico, Victor Henrique, Fábio Ferreira e Thiago Bariori. Agradecimento especial ao meu amigo jornalista Márcio Donizete, que além de ser uma grande pessoa, também é um grande profissional. Agradeço aos meus amigos torcedores do San Antonio Spurs, meu time na NBA[7] e companheiros de *fantasy* Diego Sacramento, Rafael Antunes, Vinícius Nordi Esperança (meu mestre), Vitor Lucas, Gilvan Alves, Matheus Batistussi, Vinícius Perestrelo, Pedro Henrique, João Chieregatti, Danilo Esteves, Igor Vaz Soares, Enrico Carriço, Nuno Alvim, Lucas Dahmer, Samuel Motta, Clairton Lopes e todo o grupo pelas brincadeiras, conversas e irmandades criadas nesse período. Vocês são sensacionais, obrigado!

Por fim, agradeço a base de tudo isso que conquistei e que se Ele quiser, eu ainda irei conquistar. Obrigado meu Deus, por toda essa história que o Senhor escreveu pra mim

[7] *National Association Basketball*

e que tenho fé que será escrita da melhor maneira possível. Obrigado pela minha família, meus amigos, as pessoas que o Senhor colocou na minha vida e obrigado por eu ter me tornado uma pessoa boa e por permitir que eu me torne o que eu sempre sonhei. Que o Senhor continue abençoando todos nós, amém.

"Some people wait a lifetime
for a moment like this."
(Kelly Clarkson)

Resumo

A evolução do *e-commerce* e das Redes Sociais Online (RSO) contribuiu para o aumento das informações disponíveis, tornando a tarefa de analisar comentários de forma manual praticamente impossível para o processo de tomada de decisão sobre a aquisição ou não de um produto ou serviço. Devido ao volume de informações tornou-se necessário criar métodos automáticos de extração de conhecimento. A mineração de opinião é um dos temas tratados pela comunidade de Processamento de Linguagem Natural (PLN). Atualmente, para facilitar a análise de comentários alguns sites utilizam filtros tais como, votos de utilidade ou número de estrelas. Porém, o uso desses filtros não é uma boa prática pois eles podem excluir comentários que tenham sido recentemente submetidos ao processo de votação, além de existir a possibilidade do usuário superestimar ou subestimar o comentário com a atribuição das estrelas. Uma possível solução para tais problemas é filtrar os comentários baseados na descrição textual, nas informações do autor e em outras medidas. Sousa (2015) propôs uma abordagem, denominada TOP(X), para estimar o grau de importância de comentários sobre produtos e serviços utilizando um Sistema *Fuzzy* com três variáveis de entrada: reputação do autor, extração de tuplas <característica, palavra opinativa> e analisador de riqueza e uma variável de saída: grau de importância do comentário. Apesar da abordagem apresentar bons resultados, alguns problemas ficaram pendentes de resolução e melhorias, além da possibilidade de alterar o modelo computacional utilizado. Esta Dissertação propõe adaptações em duas variáveis de entrada, a saber: quantidade de tuplas e riqueza do vocabulário e a construção de novas abordagens utilizando modelos computacionais baseados em Sistemas *Fuzzy* e Redes Neurais Artificiais (RNA). Adicionalmente, fez-se uma comparação entre as abordagens propostas por meio de medidas estatísticas. Experimentos realizados no domínio de hotéis mostraram que a abordagem utilizando Sistema *Fuzzy* obteve melhores resultados na detecção dos comentários mais importantes, sem considerar a orientação semântica dos comentários. Entretanto, a abordagem usando RNA do tipo *Multi-Layer Perceptron* (MLP) obteve melhores resultados quando se conhece a orientação semântica do comentário (positivo ou negativo).

Palavras-chaves: Mineração de Opinião. Processamento de Linguagem Natural. Redes Neurais Artificiais. Sistemas *Fuzzy*.

Abstract

The evolution of e-commerce and On-line Social Networks has contributed to the increase of the information available, making the task of analyzing the reviews manually almost impossible for the buying (or not) a product or service decision-making process. Due to the amount of information, the creation of automatic methods of knowledge extraction and data mining has become necessary. The opinion mining is one of the topics addressed by the Natural Language Processing (NLP) community. Currently, to facilitate the analysis of reviews some websites use filters such as votes by utility or by stars. However, the use of these filters is not a good practice because they may exclude reviews that have recently been submitted to the voting process, besides the possibility of the user overestimate or underestimate the review with attribution of stars. One possible solution is to filter the reviews based on their textual descriptions, author informations and others measures. Sousa (2015) proposed an approach, called TOP(X), to estimate the degree of importance of reviews about products and services using a Fuzzy System with three input variables: author reputation, extraction of tuples <feature, opinion word> and richness analyzer and an output variable: degree of importance of the review. Although the approach presented good results, some problems were pending of resolution and improvements, besides the possibility to change the computational model used. This Dissertation proposes adaptations in two input variables, namely: quantity of tuples and vocabulary richness and the building of new approaches using computational models based on Fuzzy Systems and Artificial Neural Networks (ANN). In addition, a comparison was made among the proposed approaches through statistical measures. Experiments performed in the hotel-domain showed that the approach using Fuzzy System obtained better results when detecting the most important reviews, without considering the semantic orientation of the comments. However, the approach using Multi-Layer Perceptron (MLP) Artificial Neural Networks obtained better results when is known the semantic orientation of the review (positive or negative).

Keywords: Artificial Neural Networks. Fuzzy Systems. Natural Language Processing. Opinion Mining.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADA	Árvores de Decisão Aleatórias
BM	Bom
BOW	*Bag of Words*
CIaaS	*Computational Intelligence as a Service*
CTTR	*Corrected Type-Token Ratio*
EUA	Estados Unidos da América
EXC	Excelente
FCL	*Fuzzy Controller Logic*
FN	*False Negative*
FP	*False Positive*
IA	Inteligência Artificial
IC	Inteligência Computacional
ISF	Insuficiente
LARA	*Latent Aspect Rating Analysis*
MLP	*Multi-Layer Perceptron*
MTLD	*Measure of Textual Lexical Diversity*
NRBF	*Normalized Radial Basis Function*
NLTK	*Natural Language ToolKit*
PLN	Processamento de Linguagem Natural
POS	*Part-Of-Speech*
PPGCC	Programa de Pós-Graduação em Ciência da Computação
RBF	*Radial Basis Function*
RFE	*Recursive Feature Elimination*

RNA	Redes Neurais Artificiais
RSO	Redes Sociais Online
SF	Suficiente
SPSS	*Statistical Package for the Social Sciences*
SQL	*Structured Query Language*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency–Inverse Document Frequency*
TN	*True Negative*
TP	*True Positive*
TTR	*Type-Token Ratio*
UFPI	Universidade Federal do Piauí
WOM	*Word-Of-Mouth*

Lista de símbolos

φ	Coeficiente de Correlação de Matthews
Ω	Conjunto de neurônios
γ	Conjunto das conexões entre os neurônios
Λ	Conjunto do universo de discurso U
c	Característica
F	Medida-F
k	Importância do comentário
n	Número de palavras do comentário
NC	Tamanho da amostra aleatória para treinamento e teste da rede neural artificial
P	Precisão
q	Quantidade mínima de comentários pertencente à amostra
qc	Quantidade de conjuntos *Fuzzy* a que o elemento pertence
R	Cobertura
s	Número de segmentos completos do comentário
SP	Número de segmentos parciais do comentário
t	Número de termos únicos do comentário
T	Tamanho do Córpus
u	Elemento do conjunto Λ pertencente ao universo U
U	Universo de discurso dos conjuntos clássicos e *Fuzzy*
X	Número de comentários mais importantes

Sumário

Introdução

Contexto e Motivação

Na sociedade atual, a informação é o elemento chave para qualquer organização, pois a organização que não tem informação para subsidiar suas decisões estratégicas, bem como a sua gestão, fica em desvantagem em relação às outras (LOUSADA; VALENTIM, 2011). Ter o conhecimento de dados, tais como resgatar o que a imprensa e as mídias sociais falam sobre a organização, permite a tomada de decisões mais rapidamente, com mais consistência, objetividade e precisão. Esse conhecimento tem contribuído para o surgimento de novos paradigmas de gestão empresarial e tem provocado grandes impactos sociais. Nesta sociedade da informação, os consumidores passam a fazer parte do funcionamento da empresa, na qual a qualidade dos produtos e o atendimento aos clientes são de suma importância para a sobrevivência das organizações, que são dependentes totais dos seus sistemas e tecnologias de informação (WARD; PEPPARD, 2016).

A *Web* com suas capacidades intrínsecas de compartilhamento de informações, desde seu início ganhou notoriedade mundial. Essas capacidades ganharam mais força com a evolução da *Web* primária, chamada de *Web* 2.0. Os serviços criados e popularizados pela *Web* 2.0 fizeram com que a quantidade de dados existentes na *Web* aumentasse enormemente. De acordo com Jurafsky e Martin (2008), a imensa quantidade de dados na *Web* e nas redes sociais tem tornado possível construir novas aplicações, tais como: i) perguntas e respostas (do inglês: *question answering*) (KWOK; ETZIONI; WELD, 2001; KUCUKTUNC et al., 2012); ii) extração de informação (do inglês: *information extraction*) (BOLLEGALA; MATSUO; ISHIZUKA, 2011; CHAMLERTWAT et al., 2012); iii) análise de sentimentos (do inglês: *sentiment analysis*) (FELDMAN, 2013); iv) tradução automática (do inglês: *machine translation*) (NAVIGLI; PONZETTO, 2012; VASWANI et al., 2013); v) sumarização de opiniões e documentos (do inglês: *opinion summarization*) (ZHANG; HE, 2015; FELIPPO; TOSTA; PARDO, 2016), entre outros.

Segundo Liu (2010), a informação proveniente de um texto pode ser categorizada em dois tipos principais: fatos e opiniões. Um fato é algo que aconteceu na realidade e que é de conhecimento de todos. Uma opinião é uma interpretação dos fatos, que varia de autor para autor. Liu (2010) ainda define fatos como expressões objetivas e opiniões como expressões subjetivas. Os fatos não podem ser alterados, uma vez que podem ser provados por meio de documentos. As opiniões, por serem subjetivas, divergem de acordo com o sentimento das pessoas que as emitem.

Nas atividades comerciais, quando uma pessoa tem interesse por um produto ou

serviço é comum, para a tomada de decisão de adquirir ou não o produto ou serviço, que ela procure referências ou opiniões de outras pessoas. Isto não é verdadeiro apenas para uma pessoa, mas também para organizações, uma vez que empresas que vendem produtos e disponibilizam serviços também são motivadas a ter conhecimento das opiniões das pessoas, tendo que procurar formas de analisar essas informações para conduzir ações de *marketing* e tomada de decisão. Dessa forma, opiniões dos consumidores são importantes para o sucesso ou falha de um produto ou serviço. Atualmente, os *sites* de compra e venda possuem uma seção onde seus clientes podem deixar comentários sobre o produto anunciado.

Além disso, com a evolução da *Web*, o modo com que as pessoas expressam suas opiniões também mudou. Existem vários locais como fóruns, grupos de discussão, blogs, Redes Sociais *Online* (RSO) e *sites* de compra e venda de produtos onde as pessoas escrevem textos opinativos sobre produtos e serviços, que ficam disponíveis para as outras pessoas que visitam esses locais em busca de opiniões para tomada de decisão, tornando esses locais uma grande fonte de informação com várias aplicações práticas (BONCHI et al., 2011; GIL DE ZÚÑIGA, 2012; MILNE; WITTEN, 2013).

Um dos principais locais de informação é o *e-commerce*, que engloba *sites* de compra e venda de produtos e prestação de serviços. Ultrapassando a marca de 12 milhões de lojas ao redor do planeta, o *e-commerce* é uma das principais atividades presentes na internet[8]. Devido ao grande número de páginas de produtos e serviços disponíveis, a utilização do *e-commerce* pode se tornar desgastante pois durante o processo de busca de informação, consumidores são geralmente submetidos a múltiplas escolhas (VOHS et al., 2008; BARBOSA; MOURA; SANTOS, 2016).

Outro processo de tomada de decisão de consumidores foi demonstrado por Goldenberg, Libai e Muller (2001) referente à comunicação boca-a-boca (do inglês *word-of-mouth* - WOM). Eles afirmam que esse processo é fortemente influenciado pelo WOM no qual um consumidor satisfeito certamente faria propaganda positiva sobre o produto que adquiriu para pessoas próximas, enquanto um consumidor não satisfeito faria uma propaganda negativa, afastando novos consumidores. Segundo Kim e Srivastava (2007), as RSOs concentram a comunicação WOM pois permitem que consumidores compartilhem suas experiências e recomendações por meio de opiniões, sejam escrevendo, votando, comentando ou compartilhando.

Pesquisadores da área de Processamento de Linguagem Natural (PLN) têm buscado extrair informações úteis de dados não estruturados, pois cerca de 95% das informações relevantes são originadas de forma não-estruturada, principalmente os textos tais como

[8] Disponível em https://www.internetretailer.com/commentary/2014/12/04/how-many-online-stores-are-there-world?p=1

e-mails, pesquisas, *posts* em redes sociais e fóruns, entre outros[9]. Todos os dias são criados 2,5 quintilhões de *bytes* de dados, tanto que 90% dos dados do mundo hoje foram criados apenas nos últimos dois anos[10]. Essa grande quantidade de dados faz com que a análise manual se torne uma tarefa impossível, sendo necessária a criação de métodos automáticos para analisar os dados (LIU, 2010).

A partir do advento dos sistemas computacionais, um dos principais desejos das organizações tem sido o de armazenar dados. Nos últimos tempos, com o barateamento da estrutura para armazenar uma maior quantidade de dados, essa tendência ficou ainda mais evidente. A manipulação desses dados é tratada pela área de mineração de dados (do inglês: *Data Mining*) (PIATETSKY-SHAPIRO, 1991). A mineração de dados, segundo Berry e Linoff (1997) é a exploração e a análise, por meio automático ou semiautomático, de uma grande quantidade de dados, a fim de descobrir padrões e regras significativos.

A mineração de dados é utilizada para realizar a revisão de opiniões de produtos e serviços nos mais diversos *sites* de *e-commerce* e RSOs. Várias pesquisas foram conduzidas nesse tema, que englobam técnicas para detecção e sumarização automáticas de opinião sobre revisões de produtos e serviços (DAVE et al., 2003; HU; LIU, 2004; ARCHAK; GHOSE; IPEIROTIS, 2007), focam sobre entidades específicas (por exemplo políticos, celebridades, marcas) em redes sociais (GUERRA et al., 2011) ou notícias (GODBOLE; SRINIVASAIAH; SKIENA, 2007), realizam mineração de opinião sobre textos menos estruturados, como notícias e blogs (KU; LIANG; CHEN, 2006; BALAHUR et al., 2009) e analisam a opinião do *Twitter*[11] para estabelecer modelos preditivos (GHANI et al., 2006; ASUR; HUBERMAN, 2010).

Um dos domínios existentes no *e-commerce* utilizado pela mineração de dados é o de hotéis, onde de acordo com Kasper e Vela (2011), o planejamento e reserva de viagens se tornou um dos mais importantes tópicos no comércio, principalmente devido às reservas de hotéis. O site *Booking.com* é um dos sites mais utilizados para este fim no Brasil, uma vez que pesquisa realizada em 2014, o *Booking.com* foi escolhido o melhor site de reserva de hotéis na avaliação de leitores[12]. O *Booking.com* é um *website* de reserva de hotéis disponível em 40 idiomas e ativo em 226 países e territórios[13]. No Brasil, o *Booking.com* tem 32.542 hotéis cadastrados[14]. De acordo com a página inicial do site, apenas hóspedes podem escrever avaliações. Em um teste realizado em uma das páginas do hotel, foi verificado que realmente não há local para avaliações de usuários que não se hospedaram em algum hotel. Apenas no perfil do hóspede há as opções de avaliações.

[9] Disponível em http://www.clarabridge.com/nlp-natural-language-processing/
[10] http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf
[11] http://twitter.com
[12] Disponível em http://www.melhoresdestinos.com.br/melhor-site-reservar-hoteis.html
[13] Disponível em http://www.booking.com/content/about.pt-br.html
[14] Disponível em https://www.booking.com/country.pt-br.html. Dados de dezembro/2016

Como existe um grande número de comentários publicados pelos usuários, as opiniões são classificadas normalmente por estrela, pelas mais recentes ou mais relevantes, porém nem sempre são as opiniões mais importantes ou úteis para um determinado usuário. Em alguns sites de hotéis, usuários podem votar em comentários que eles consideram úteis ou inúteis quando estão pesquisando sobre o hotel. Entretanto, nem sempre apenas informações de polaridade do comentário são suficientes pois outros problemas podem acontecer, como destacado por Li et al. (2013): comentários mais novos que ainda não foram votados serão dificilmente lidos e votados. Assim, disponibilizar os comentários mais importantes, baseados na descrição textual, na riqueza do vocabulário e na qualidade do autor é um fator que deve ser considerado. Dessa forma, novos hóspedes podem analisar um pequeno conjunto de comentários para a tomada de decisão.

Objetivos

Com o objetivo de identificar os comentários mais relevantes, Sousa (2015) propôs a abordagem TOP(X) para inferir os X melhores comentários sobre produtos ou serviços. A abordagem TOP(X) utiliza um Sistema *Fuzzy* com três variáveis de entrada: reputação do autor, extração de tuplas <característica, palavra opinativa> e analisador de riqueza; e uma variável de saída: grau de importância do comentário, representado pela variável k (ver Figura 1). Uma alternativa a abordagem TOP(X) é substituir o Sistema *Fuzzy* por uma Rede Neural Artificial, pois os conceitos da Lógica *Fuzzy* podem ser utilizados em áreas de aplicação das Redes Neurais Artificiais, tais como classificação de padrões e otimização de sistemas. (TANSCHEIT, 2004; HAYKIN, 2009).

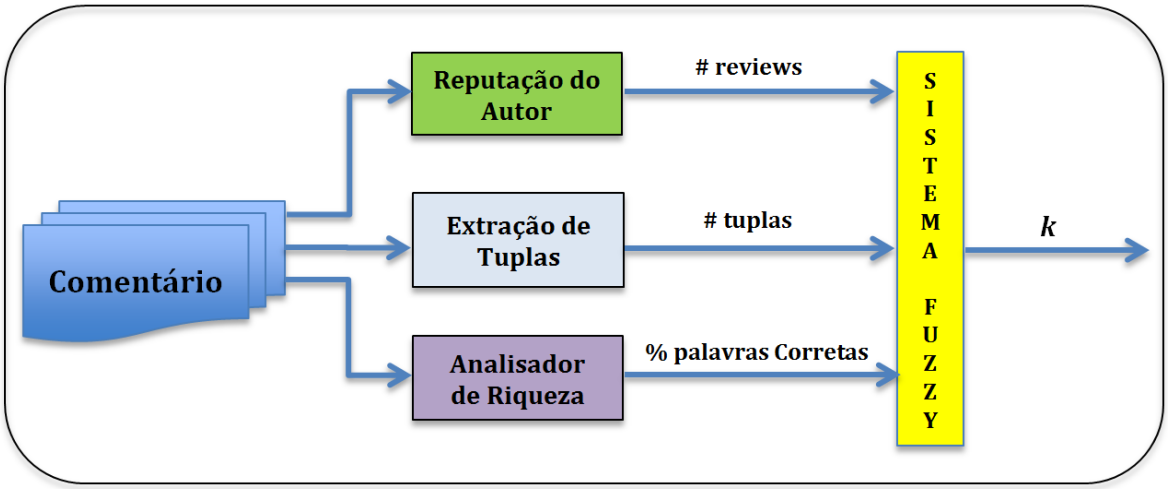


Figura 1 – Estrutura geral do Sistema *Fuzzy* por Sousa (2015).

Assim, o objetivo principal deste trabalho consiste em propor adaptações em duas variáveis de entrada da abordagem TOP(X), a saber: **extração de tuplas** e **analisador de riqueza**, denominadas, nesta Dissertação, de **quantidade de tuplas** e **riqueza do**

vocabulário. Além disso, este trabalho propõe abordagens utilizando Sistemas *Fuzzy* e Redes Neurais Artificiais para fazer uma comparação por meio de medidas estatísticas. Adaptações na variável reputação do autor está fora do escopo deste trabalho, pois elas tem sido exploradas por outro pesquisador do grupo de pesquisa.

São considerados objetivos específicos deste trabalho:

- Criar um Córpus anotado manualmente, a fim de realizar os testes com as abordagens propostas;
- Construir uma base de características e propor novos padrões linguísticos para extrair as tuplas <característica, palavra opinativa> no domínio de hotéis;
- Utilizar novas medidas de riqueza léxica do texto para aumentar a confiabilidade da variável referente à riqueza do vocabulário;
- Modelar abordagens utilizando Sistemas *Fuzzy* e Redes Neurais Artificiais a partir das adaptações propostas para estimar o grau de importância de comentários Web;
- Realizar experimentos com as abordagens propostas para estimar o grau de importância dos comentários a fim de definir qual abordagem obtém melhores resultados.

Vale mencionar que este trabalho faz parte de um projeto maior que está sendo desenvolvido no PPGCC-UFPI, que realiza a extração de características de um produto em sites de fabricantes, analisa os comentários de usuários em sites de compra, venda ou comparação de preços de produtos e analisa também as reclamações realizadas sobre o produto ou empresa no site Reclame Aqui[15].

Contribuições

Destaca-se como principais contribuições:

- As adaptações no modelo com Sistema *Fuzzy*, as quais têm como objetivo uma acurácia maior na estimação do grau de importância do comentário. Além disso, essas adaptações englobam outras áreas do conhecimento, como por exemplo a utilização de índices de medida do vocabulário, usado na área de Linguística;
- O modelo computacional utilizando Redes Neurais Artificiais, o qual abrange a ideia previamente proposta para outros modelos computacionais semelhantes;

[15] http://www.reclameaqui.com.br/

- A comparação entre os modelos computacionais, na qual pode direcionar o objetivo de estimar o grau de importância para outras linhas de pesquisa em inteligência artificial;

- O Córpus[16] anotado, o qual possibilita a outros pesquisadores como base para trabalhos futuros ou replicação deste trabalho.

Partes desta pesquisa foram publicados na comunidade acadêmica em diferentes conferências. Tais artigos são descritos a seguir:

- *An Experimental Study based on Fuzzy Systems and Artificial Neural Networks to Estimate the Importance of Reviews about Product and Services* (SANTOS et al., 2016a): Este trabalho apresenta a primeira versão deste trabalho proposto, sendo um estudo experimental baseado em abordagens utilizando Sistema *Fuzzy* e Redes Neurais Artificiais. A comparação entre as abordagens foi realizada tendo como base o trabalho de Sousa (2015) com uma abordagem utilizando RNA com as mesmas variáveis de entrada no mesmo domínio de aplicação. Nesse trabalho, a abordagem com Sistema *Fuzzy* de Sousa (2015) se mostrou superior na detecção dos comentários mais importantes. Vale ressaltar que não foi alterada a maneira como as variáveis de entrada foram obtidas, diferentemente do que está sendo proposto nesta Dissertação.

- *Extração de Métricas e Análise de Sentimentos em Comentários Web no Domínio de Hotéis* (SANTOS; MOURA, 2016): Este trabalho apresenta um protótipo de uma aplicação que sumariza informações de um comentário tais como o sentimento, suas características encontradas e outras métricas de análise. A metodologia desse trabalho foi a versão preliminar das adaptações realizadas na variável quantidade de tuplas proposta nesta Dissertação. Em relação aos resultados, a aplicação teve precisão de comentários com polaridade negativa e cobertura de comentários positivos em 84.93% e 94.33% respectivamente.

- *Evaluating the Importance of Web Comments Through Metrics Extraction and Opinion Mining* (SANTOS et al., 2016b): Este trabalho apresenta uma parte da segunda versão do trabalho proposto nesta Dissertação, especificando a abordagem com Sistema *Fuzzy* já com as adaptações nas variáveis referentes à quantidade de tuplas e uma versão inicial da adaptação na variável referente à riqueza do vocabulário, utilizando mais dois índices de riqueza léxica: CTTR (CARROLL, 1964) e Maas, se juntando à corretude do comentário. Ao executar a abordagem com Sistema *Fuzzy* proposta nesse trabalho, foram obtidos resultados que chegaram a 50% na medida-F na detecção dos comentários positivos e negativos.

[16] Foi utilizada esta nomenclatura para indicar a coleção de documentos (comentários), que indicará os termos *corpus* e *corpora* usados por outros pesquisadores.

Organização

Contando com este capítulo de Introdução, outros cinco capítulos estão presente nesta Dissertação, além das referências bibliográficas, apêndices e anexos. A organização dos capítulos está conforme detalhado a seguir:

No Capítulo 1 - Referencial Teórico, são apresentadas as áreas de pesquisa deste trabalho, tais como o Processamento de Linguagem Natural, enfatizando Mineração de Opinião ou Análise de Sentimentos e Aprendizado de Máquina, incluindo modelos computacionais baseados em Sistemas *Fuzzy* e as Redes Neurais Artificiais. No final são discutidas as ferramentas e recursos utilizados durante a realização do trabalho: etiquetadores, léxicos e o ambiente de teste para utilização dos sistemas de Inteligência Artificial, além do Córpus de referência usado para avaliação das abordagens.

No Capítulo 2 - Trabalhos Relacionados, apresenta-se uma revisão da literatura englobando os principais trabalhos da área de mineração de opinião em cada uma de suas etapas, além da utilização de algoritmos de aprendizagem de máquina em pesquisas referentes ao PLN.

No Capítulo 3 - Abordagem com Sistema *Fuzzy* e Rede Neural Artificial, descreve-se o detalhamento das abordagens com os modelos computacionais baseados em Sistemas *Fuzzy* e Redes Neurais Artificiais, além das adaptações propostas nas variáveis referentes à quantidade de tuplas e riqueza do vocabulário.

No Capítulo 4 - Experimentos e Discussões, são descritos os experimentos realizados em cada uma das abordagens propostas nesta Dissertação: com Sistema *Fuzzy* e as abordagens com Redes Neurais Artificiais, detalhes de cada execução e a comparação entre as abordagens.

Por fim, no último capítulo apresenta-se as conclusões, limitações e os trabalhos futuros propostos para continuação da pesquisa.

1 Referencial Teórico

Neste capítulo são apresentadas as áreas de interesse desta Dissertação, destacando os principais conceitos e definições referentes à Mineração de Opinião, bem como os algoritmos de aprendizagem de máquina utilizados, tais como Sistemas *Fuzzy* e Redes Neurais Artificiais. Ao final, destaca-se as ferramentas e recursos utilizados neste trabalho.

1.1 Processamento de Linguagem Natural

O termo "Processamento de Linguagem Natural" (PLN) é geralmente utilizado para descrever a função de componentes de *software* ou *hardware* em um sistema computacional que analisa ou sintetiza uma linguagem falada ou escrita (JACKSON; MOULINIER, 2002). Bird, Klein e Loper (2012) definem como "linguagem natural" uma linguagem que é usada para comunicação diária por humanos como inglês, hindi ou português. O termo "natural" é usado para diferenciar a fala e a escrita humana das linguagens mais formais, por exemplo, notações lógicas ou matemáticas, ou linguagens de programação tais como Java e C++ (JACKSON; MOULINIER, 2002).

Tecnologias baseadas em PLN estão se tornando cada vez mais difundidas. Atualmente existem *smartphones* que suportam previsão de textos e reconhecimento de voz, além dos mecanismos de pesquisa *Web* que dão acesso às informações encontradas em textos não-estruturados e tradutores que permitem apresentar textos escritos em uma linguagem em uma outra em questão de segundos. Jackson e Moulinier (2002) afirmam que o processamento da linguagem tem uma importante função tanto na produção quanto na sumarização da informação *online*.

A área de PLN contém uma importante subárea: a extração de informação, que é caracterizada por duas propriedades: i) o conhecimento desejado pode ser descrito por um modelo relativamente simples com lacunas que necessitam ser preenchidas com material do texto; e ii) apenas uma pequena parte da informação no texto é relevante para preencher essa lacuna, onde o resto pode ser ignorado (JURAFSKY; MARTIN, 2008). Portanto, programas que utilizam extração de informação analisam apenas um pequeno subconjunto de um dado texto, por exemplo, aquelas partes que contêm certas palavras que representam os objetos ou eventos de interesse (JACKSON; MOULINIER, 2002).

Junto à extração de informação, técnicas de PLN buscam extrair uma representação do significado de textos livres, ou seja, descobrir em um texto elementos utilizados para organizar os pensamentos: quem, o que, como, onde e por quê (ROBERTSON, 1946). Esses elementos são extraídos utilizando conceitos linguísticos como classes gramaticais

(substantivos, verbos, adjetivos e advérbios) e estruturas gramaticais (sintagmas nominais, verbais e preposicionais), bem como o tratamento de desafios como a resolução de anáforas, ambiguidades e erros provenientes da forma informal que atualmente são escritos os textos na *Web*, como por exemplo, abreviações, gírias ou neologismos.

Outra subárea fundamental de PLN é a Mineração de Opinião que consiste no estudo computacional de opiniões, avaliações, atitudes e emoções das pessoas referenciando entidades, indivíduos, questões, eventos, tópicos e seus atributos. O estudo é realizado sobre a emoção do usuário em relação ao item, classificando a emoção em positiva, negativa ou neutra. Mais detalhes são discutidos a seguir.

1.1.1 Mineração de Opinião

De forma científica, Liu (2010) define uma opinião como sendo uma quíntupla (e_i, a_{ij}, s_{ijkl}, h_k, t_l) onde:

- e_i é o nome de uma entidade;
- a_{ij} é um aspecto de uma entidade;
- s_{ijkl} é o sentimento sobre um aspecto a_{ij} da entidade e_i, emitido pela pessoa h_k no tempo t_l;
- h_k é o detentor da opinião (*opinion holder*);
- t_l é o instante no qual a opinião foi expressada por h_k.

Algumas observações para essa definição são importantes (LIU, 2010). Todos os componentes da quíntupla são essenciais, uma vez que a falta de um deles traria problemas em geral, por exemplo, o aspecto "tela" pode se referir à várias outras entidades como *smartphones* ou câmeras. Em complemento, todos os componentes da quíntupla devem corresponder um ao outro, por exemplo, o sentimento s_{ijkl} deve ser determinado pelo detentor da opinião h_k sobre o aspecto a_{ij} da entidade e_i no instante t_l, no qual qualquer incompatibilidade torna-se um erro. Outra observação é que a definição cobre a maioria das opiniões, mas não todos os aspectos do significado semântico da opinião, que pode ser complexo. O exemplo dado é que a frase "Este carro é muito pequeno para uma pessoa alta" não diz que o carro é pequeno para todos, sendo que "pessoa alta" é o contexto. Por fim, a definição dada é válida para um tipo de opinião chamada opinião regular, diferente da opinião comparativa que necessita de uma definição diferente (JINDAL; LIU, 2006).

O conceito de aspecto, também denominado característica (*feature*) ou propriedade, permite que uma entidade seja vista por meio de diferentes perspectivas ou atributos, ou como uma hierarquia de partes e subpartes. Zhang e Liu (2014) além de conceituarem

aspecto, também definiram expressão de aspecto como sendo uma palavra atual ou uma frase que aparece no texto indicando um aspecto. Como exemplo, os autores utilizam o domínio de celulares e exemplificam o aspecto "qualidade de voz" e discutem que várias expressões como "som" e "voz" podem indicar o aspecto alvo.

A detecção do sentimento em um texto pode ocorrer em diferentes granularidades, sendo que a decisão do nível de análise textual está sujeita a seu contexto e aplicação. Essa análise, de acordo com Liu (2012), pode ser no nível de:

- **Documento**: a tarefa principal é classificar se um documento tratado como um todo expressa o sentimento positivo ou negativo. Um exemplo de utilização desse nível é quando o documento trata de uma única entidade, como por exemplo, uma opinião sobre um dado evento;
- **Sentença**: o objetivo é determinar o sentimento de uma sentença específica de um certo documento. Um exemplo de utilização desse nível é quando o documento contém opiniões sobre várias entidades;
- **Entidade ou Aspecto**: busca-se focar na opinião expressa. A utilização desse nível se dá quando o alvo da opinião possa ser uma entidade ou algum dos seus aspectos. É o nível mais complexo de análise.

Dos níveis de análise textual apresentados, o nível de aspectos é o mais desafiador, uma vez que contém vários problemas de PLN ainda não resolvidos como o reconhecimento de entidades nomeadas, resolução de anáfora, negação, ironias e sarcasmos (LIU, 2010).

Além disso, opiniões referem-se a um conteúdo subjetivo, escrito na linguagem natural. Assim, a forma como essas opiniões estão expressas tem influência na capacidade de processá-las de forma correta. A mineração de opiniões tem origens em comum com a linguística computacional, com a qual compartilha problemas e desafios (LIU, 2012).

Liu (2012) ainda define os tipos de opiniões como sendo:

- **Regulares ou Comparativas**: Uma opinião é regular quando o autor da opinião expressa um sentimento, atitude, emoção ou percepção sobre um alvo. As opiniões comparativas expressam o sentimento baseadas na relação de similaridades ou diferenças entre duas ou mais entidades ou quando há algum aspecto compartilhado. O Exemplo 1 mostra uma opinião regular. O autor da opinião fala diretamente sobre o aspecto do hotel exprimindo o sentimento em "Atendimento excelente", "ótima localização" e " café da manhã completíssimo". Uma opinião comparativa é apresentada no Exemplo 2 a seguir, onde o autor descreve em "Fiquei em hotéis com menos nomes e com estrelas inferiores muito melhor" a comparação com o hotel em destaque na opinião com outros existentes;

Exemplo 1 "Atendimento excelente, profissionais muito educados e disposto, ótima localização, café da manhã completíssimo, quarto limpo, ar condicionado potente, cama confortável, elevadores ágeis."

Exemplo 2 "Por se tratar de um hotel 5 estrelas eu esperava mais das acomodações do quarto. Fiquei em hotéis com menos nomes e com estrelas inferiores muito melhor. Além disso, deixei bem claro que queria uma cama de casal, fato que não ocorreu."

Em destaque, uma opinião regular é definida na literatura apenas como "opinião" e tem mais dois subtipos (LIU, 2006):

- **Diretas ou Indiretas**: Identifica-se uma opinião direta quando há a referência direta a uma entidade ou um aspecto da entidade. No caso das opiniões indiretas, a opinião é expressa de forma indireta sobre uma entidade ou aspecto da entidade. Na opinião do Exemplo 3 pode ser observada a referência direta a hotel no trecho "O hotel é muito bom". As opiniões indiretas ocorrem com mais frequência no domínio médico. Na opinião do Exemplo 4, é descrito um efeito indesejado do remédio em relação à gripe, o que indiretamente relaciona uma opinião negativa ao sentimento do remédio;

Exemplo 3 "O hotel é muito bom, a localização é maravilhosa. Copacabana é tudo de bom, café da manhã é excelente, pessoal muito prestativo e atenciosos. Gostamos de tudo!!!!!"

Exemplo 4 "Minha gripe piorou depois que tomei este remédio."

- **Explícitas ou Implícitas**: Nas opiniões explícitas o sentimento é exposto de forma direta enquanto que nas opiniões implícitas o sentimento é expressado indiretamente. Um exemplo de opinião com os dois tipos citados é mostrado no Exemplo 5. No trecho "Café da manhã muito bom" o autor expressa sua opinião positiva explicitamente sobre o aspecto "café da manhã". O mesmo autor exprime o sentimento sobre o aspecto comida de forma explícita (usando o termo "excelente") e de forma implícita, usando o termo "cara", referenciando o preço alto do aspecto.

Exemplo 5 "Café da manhã muito bom. Excelente tratamento nos bares e restaurante. Comida excelente e não é cara."

É importante destacar que a maioria dos trabalhos concentram-se sobre as opiniões regulares, diretas e explícitas, devido a facilidade de detecção de tais tipos de opiniões (BECKER; TUMITAN, 2013).

De acordo com Tsytsarau e Palpanas (2012), a mineração de opinião pode ser caracterizada em termos de três grandes tarefas: i) identificar (tópicos, sentenças opinativas);

ii) classificar a polaridade do sentimento; e iii) sumarizar. A Figura 2 mostra uma visão geral do processo de extração de características de comentários *Web* e as três etapas de mineração de opinião.

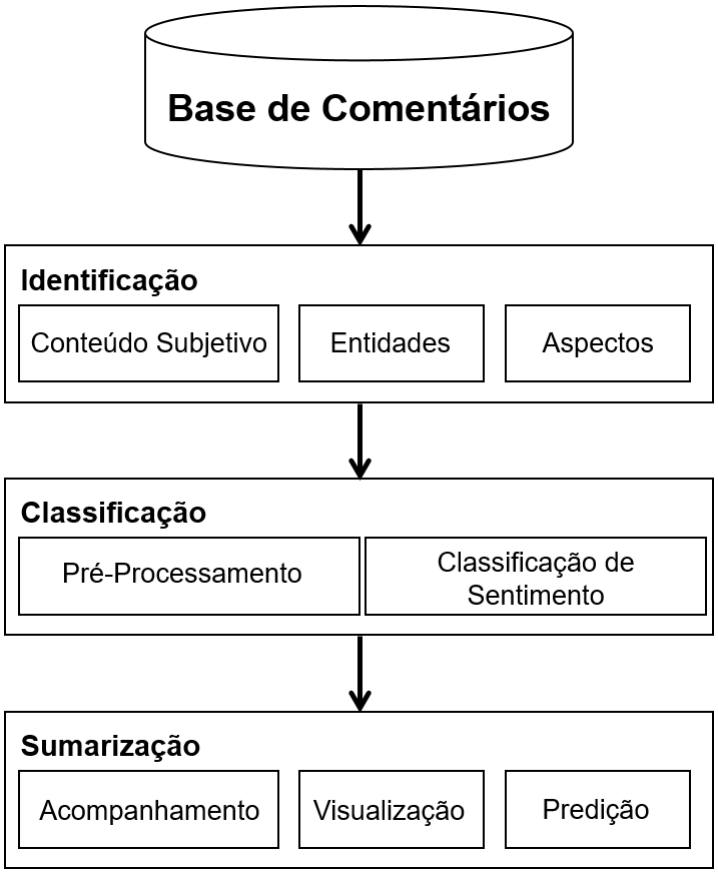


Figura 2 – Etapas da Mineração de Opinião (BECKER; TUMITAN, 2013) (adaptado)

A etapa de identificação consiste em encontrar as entidades existentes (e porventura seus aspectos) e possivelmente associá-los com o respectivo conteúdo subjetivo. A etapa de classificação de polaridade é a que classifica o dado conteúdo subjetivo[1] em uma de três classes: positivo, negativo ou neutro. No entanto, podem ter outras classes adicionais com aumento do nível de detalhe dos resultados, como muito positivo, moderadamente positivo, entre outros, representando um grau de intensidade (THET; NA; KHOO, 2010). Por fim, a etapa de sumarização cria métricas que representam o sentimento geral, as quais podem ser visualizadas ou servir de entrada para outras aplicações, como por exemplo, um sumário de um determinado produto para que o usuário possa identificar seus respectivos pontos fortes e fracos, levando em consideração a experiência prévia de outras pessoas, expressas em suas opiniões.

Por fim, os tópicos estudados no âmbito de Mineração de Opinião abrangem, principalmente:

[1] Sentenças com conteúdo objetivo também têm sentimentos, como por exemplo, "*A máquina parou de trabalhar no segundo dia*", mas são relativamente raras (LIU, 2012).

- **O seu problema**: para cada problema científico é necessário definir o problema, introduzindo as definições básicas, os principais conceitos e questões, subproblemas e objetivos alvo;
- **A classificação do sentimento e subjetividade**: é a área mais pesquisada na academia, classificando o documento ou a sentença em positivo, negativo ou neutro;
- **A análise de sentimentos baseada em características**: primeiro tenta-se descobrir os alvos que a opinião tem expressa na sentença e então determina se as opiniões são positivas, negativas ou neutras;
- **A análise de sentimentos em sentenças comparativas**: a avaliação de um objeto pode ser feita de duas maneiras: na opinião direta, dando opinião positiva ou negativa sobre o objeto sem mencionar qualquer outro objeto similar e na opinião comparativa, que significa comparar o objeto com outro objeto similar.

Mais detalhes sobre definições, aplicações e outras informações sobre Mineração de Opinião podem ser encontrados em Liu (2012), Liu (2015), Pozzi et al. (2016) e Pedrycz e Chen (2016).

1.2 Aprendizado de Máquina

Simon (1983) define aprendizado como qualquer mudança num sistema que melhore o seu desempenho na próxima vez que ele repetir a mesma tarefa, ou numa outra tarefa similar. O aprendizado envolve generalização a partir da experiência: o desempenho deve melhorar não apenas na "repetição da mesma tarefa", mas também em tarefas similares no domínio.

Luger (2004) descreve que a habilidade de aprender deve fazer parte de qualquer sistema que reivindique possuir inteligência num sentido geral. Agentes inteligentes devem ser capazes de se modificarem ao longo do curso de suas interações com o mundo, bem como pela experiência de seus próprios estados e processos internos.

As habilidades mais difíceis de serem computadorizadas são a linguagem e o aprendizado (LUGER, 2004). Essa afirmação pode ser explicada pelo fato de que elas englobam muitas outras habilidades inteligentes humanas. Ao longo dos anos, estas duas áreas tem funcionado como objetivo, desafio e meio de teste para o progresso da Inteligência Artificial (IA).

A aprendizagem de máquina se mostrou uma área fértil de pesquisa, produzindo uma série de problemas tais como reconhecimento, classificação e clusterização (MITCHELL, 1997), além de algoritmos diferentes para a sua solução, tais como Redes Neurais Artificiais (SERGIENKO et al., 2015; SANTOS et al., 2016a), Sistemas *Fuzzy* (ROSS, 2004; SILER;

BUCKLEY, 2005), Computação Evolutiva (GOLDBERG, 1989; MICHALEWICZ, 1996), Inteligência Coletiva (KENNEDY; EBERHART, 2001), Sistemas Imunológicos Artificiais (DASGUPTA, 2006) e Agentes Inteligentes (D'INVERNO; LUCK, 2004). Estes algoritmos variam nos seus objetivos, na disponibilidade de dados de treinamento e nas estratégias de aprendizagem e linguagem de representação do conhecimento que eles empregam.

Nas subseções seguintes serão apresentados os conceitos, objetivos e funcionamento dos modelos computacionais utilizados neste trabalho, a saber: Sistemas *Fuzzy* e Redes Neurais Artificiais.

1.2.1 Sistemas *Fuzzy*

A Teoria de Conjuntos *Fuzzy* (ZADEH, 1965) e os conceitos de Lógica *Fuzzy* (ZADEH, 1975a; ZADEH, 1975b) podem ser utilizados para traduzir, em termos matemáticos, a informação imprecisa expressa por um conjunto de regras linguísticas. Se um operador humano for capaz de articular sua estratégia de ação como um conjunto de regras da forma <se, então> um algoritmo passível de ser implementado em computador pode ser construído. O resultado é um sistema de inferência baseado em regras, no qual a Teoria de Conjuntos *Fuzzy* e Lógica *Fuzzy* fornecem o ferramental matemático para se lidar com tais regras linguísticas (TANSCHEIT, 2004).

Os conceitos de conjuntos *Fuzzy* tem o objetivo de fornecer um ferramental matemático para o tratamento de informações de caráter impreciso ou vago. A Lógica *Fuzzy*, baseada nessa teoria, foi inicialmente construída a partir dos conceitos já estabelecidos da lógica clássica. Os conjuntos *Fuzzy* são caracterizados por suas funções de pertinência. Consequentemente, as propriedades dos conjuntos *Fuzzy* são definidas diretamente das propriedades das funções de pertinência. Características geométricas dos conjuntos *Fuzzy* ajudam a visualizar e enfatizar similaridades e diferenças entre conjuntos clássicos (conjuntos *crisp*) e conjuntos *Fuzzy* (PEDRYCZ; GOMIDE, 2007).

O conceito de pertinência de um elemento a um conjunto fica bem definido quando se tratam dos conjuntos clássicos. Dado um conjunto Λ em um universo U, os elementos deste universo simplesmente pertencem ou não pertencem àquele conjunto, o qual pode ser expresso pela função característica f_Λ:

$$f_\Lambda(u) = \begin{cases} 1, & \text{se e somente se} \quad u \in \Lambda \\ 0, & \text{se e somente se} \quad u \notin \Lambda \end{cases}$$

O conjunto *Fuzzy* proposto por Zadeh (1965) tem uma caracterização mais ampla, generalizando a função característica de forma que ela possa assumir um número infinito de valores no intervalo [0,1]. Dado um conjunto *Fuzzy* Λ no universo U é definido por uma função de pertinência $\mu_\Lambda(u) = U \rightarrow [0, 1]$ e representado por um conjunto de pares

ordenados

$$\Lambda = \{\mu_\Lambda(u)/u\}, u \in U$$

onde $\mu_\Lambda(u)$ indica o quanto u é compatível com o conjunto Λ. Um elemento pode pertencer a mais de um conjunto *Fuzzy* com diferentes graus de pertinência (TANSCHEIT, 2004)

Sistemas *Fuzzy* são utilizados em várias áreas de aplicações. A inferência *Fuzzy* é um processo de avaliação de entradas (variáveis) com o objetivo de obter conclusões utilizando-se a teoria de conjuntos *Fuzzy* por meio de regras previamente definidas e das entradas fornecidas. Um sistema de inferência contém três estágios: i) fuzzificação; ii) processo de inferência Fuzzy; e iii) defuzzificação (BAI; WANG, 2006). O modelo de um sistema de inferência *Fuzzy* é mostrado na Figura 3.

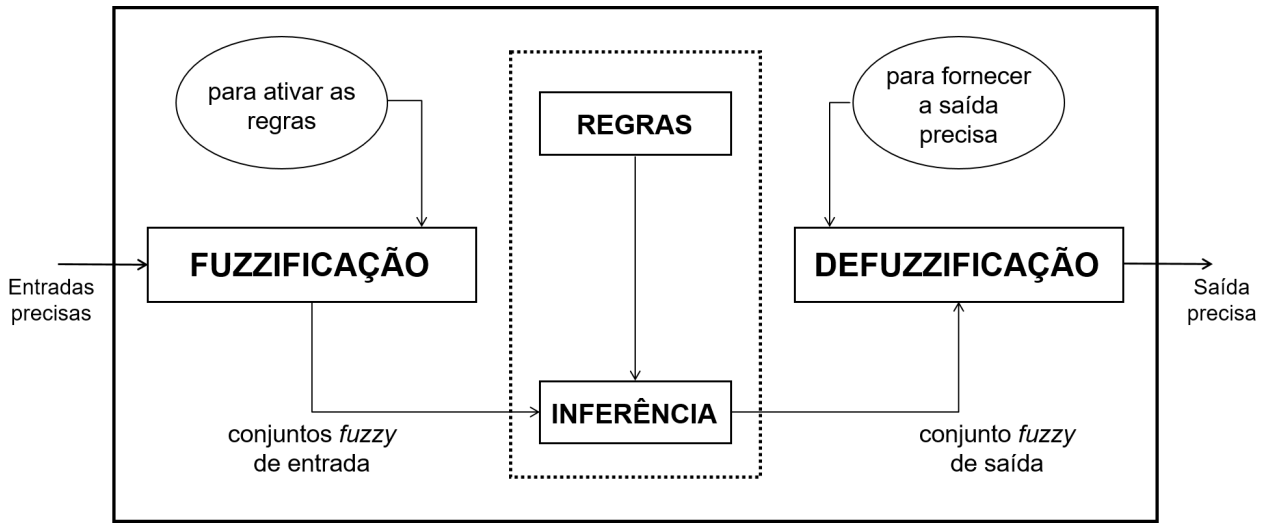


Figura 3 – Modelo de Sistema de Inferência *Fuzzy* (TANSCHEIT, 2004)

Neste sistema, consideram-se entradas não-*Fuzzy*, ou precisas, que são resultantes de medições ou observações (por exemplo, um conjunto de dados), que é o caso da grande maioria das aplicações práticas. Em virtude disto, é necessário efetuar um mapeamento destes dados precisos para os conjuntos *Fuzzy* (de entrada) relevantes, o que é realizado no estágio de fuzzificação.

As regras podem ser fornecidas por especialistas, em forma de sentenças linguísticas e se constituem em um aspecto fundamental no desempenho de um sistema de inferência *Fuzzy*. Tomando o exemplo de um controlador *Fuzzy*, este só terá um bom desempenho se as regras que definem a estratégia de controle forem consistentes. Extrair regras de especialistas na forma de sentenças do tipo <se, então> pode não ser uma tarefa fácil, por mais conhecedores que eles sejam do problema em questão. Alternativamente ao uso de especialistas para a definição da base de regras, existem métodos de extração de regras a partir de dados numéricos. Esses métodos são particularmente úteis em problemas de classificação e previsão de séries temporais (TANSCHEIT, 2004).

No estágio de inferência ocorrem as operações com conjuntos *Fuzzy* propriamente ditas: combinação dos antecedentes das regras, implicação e *modus ponens* generalizado. Os conjuntos *Fuzzy* de entrada, relativos aos antecedentes das regras, e os de saída, referentes ao consequente, podem ser definidos previamente ou, alternativamente, gerados automaticamente a partir dos dados.

Uma vez obtido o conjunto *Fuzzy* de saída por meio do processo de inferência, no estágio de defuzzificação é efetuada uma interpretação dessa informação. Este trabalho é necessário pois geralmente são requeridas saídas precisas nas aplicações práticas. Existem vários métodos de defuzzificação na literatura. Dois dos métodos mais empregados são a média dos máximos (MAMDANI; ASSILIAN, 1975) e o centro de área. Neste último método, a saída precisa é obtida tomando-se a média entre os dois elementos extremos do universo que correspondem aos maiores valores da função de pertinência do consequente. Com o centro de área, a saída é o valor do universo que divide a área sob a curva da função de pertinência em duas partes iguais.

Segundo Tanscheit (2004), um aspecto importante é a definição dos conjuntos *Fuzzy* correspondentes às variáveis de entrada (antecedentes) e às de saída (consequentes), pois o desempenho do sistema de inferência depende do número de conjuntos, de sua forma e da distribuição desses conjuntos ao longo do universo de discurso. A integração entre sistemas de inferência *Fuzzy* e redes neurais, originando os sistemas neuro-*Fuzzy* (JANG; SUN, 1997), tem se mostrado adequada para a sintonia de funções de pertinência, assim como para a geração automática de regras.

Outras informações sobre definições, tipos, uso e mais detalhes sobre Sistemas *Fuzzy* são encontrados em Cox (1994), Tanscheit (2004), Chen e Pham (2005) e Nedjah e Mourelle (2010).

1.2.2 Redes Neurais Artificiais (RNA)

Redes Neurais Artificiais são modelos computacionais inspirados no sistema nervoso de seres vivos, sendo uma tentativa de modelar as capacidades de processamento de informação dos sistemas nervosos (ROJAS, 1996). Possuem a capacidade de aquisição e manutenção do conhecimento (baseado em informações) e podem ser definidas como um conjunto de unidades de processamento, denominadas por neurônios artificiais, que são interligados por um grande número de interconexões, chamadas de sinapses artificiais, as quais são representadas por vetores ou matrizes de pesos sinápticos (SILVA; SPATTI; FLAUZINO, 2010).

A estrutura das RNAs foi desenvolvida a partir de modelos conhecidos de sistemas nervosos biológicos e do próprio cérebro humano. Os elementos computacionais ou unidades processadoras, denominadas neurônios artificiais, são modelos simplificados dos neurônios

biológicos, inspirados no trabalho de Hodgkin e Huxley (1952). Os neurônios artificiais utilizados nos modelos de RNAs não-lineares fornecem saídas tipicamente contínuas e realizam funções simples, como coletar os sinais existentes em suas entradas, agregá-los de acordo com sua função operacional e produzir uma resposta, levando em consideração sua função de ativação inerente.

O modelo proposto por McCulloch e Pitts (1943) engloba as principais características de uma rede neural biológica, sendo o modelo mais simples e o mais utilizado nas diferentes arquiteturas de RNAs. A Figura 4 mostra o modelo de neurônio artificial. O funcionamento de um neurônio artificial começa pela apresentação dos sinais de entrada representados pelo conjunto $\{x_1, x_2, ..., x_n\}$ que são ponderados por meio do conjunto de pesos sinápticos $\{w_1, w_2, ..., w_n\}$. Depois é obtido o potencial de ativação μ produzido pela soma ponderada dos sinais de entrada através do combinador linear Σ, subtraindo-se ao limiar de ativação θ. Com o potencial de ativação calculado, é aplicada a função de ativação $g(.)$ apropriada, tendo-se como objetivo limitar a saída do neurônio. Finalmente é convertida a saída y a partir da aplicação da função de ativação neural em relação ao seu potencial de ativação.

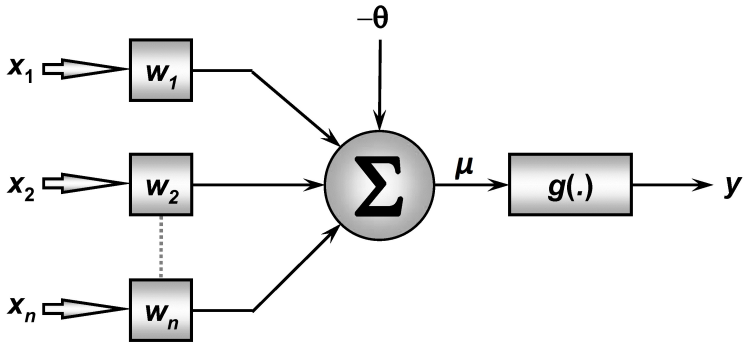


Figura 4 – Neurônio artificial (SILVA; SPATTI; FLAUZINO, 2010)

Em Kriesel (2007) é mostrada matematicamente uma rede neural artificial sendo uma tripla (Ω, γ, w) com dois conjuntos Ω, γ e uma função w, onde Ω é o conjunto de neurônios e γ é um conjunto $\{(i, j) \mid i, j \in \mathbb{N}\}$ cujo elementos são chamados de conexões entre o neurônio i e o neurônio j. A função $w : \gamma \rightarrow \mathbb{R}$ define os pesos sinápticos, onde $w((i, j))$ é o peso da conexão entre o neurônio i e o neurônio j, o qual é reduzido para w_{ij}.

Uma RNA pode ser dividida em três camadas: camada de entrada, camadas escondidas (ou intermediárias, ocultas ou invisíveis) e camada de saída. A camada de entrada é responsável pelo recebimento das informações externas. As camadas escondidas são compostas por neurônios que extraem as características associadas ao processo ou sistema inferido. Quase todo o processamento interno da rede é realizado nessas camadas. Por fim, a camada de saída é responsável pela produção e apresentação dos resultados finais da rede, resultante dos processamentos efetuados pelos neurônios das camadas anteriores.

As funções de ativação $g(.)$ são utilizadas nas camadas escondidas e camadas de

saída com o objetivo de limitar a saída do neurônio dentro de um intervalo de valores razoáveis a serem assumidos pela sua própria imagem funcional. Das funções de ativação existentes na literatura, três são importantes destacar, a saber: identidade, tangente-hiperbólica e *softmax*. A função identidade retorna sempre o mesmo valor que foi utilizado como argumento, ou seja, os resultados são idênticos aos valores do limiar de ativação μ , na qual sua expressão matemática definida por

$$g(\mu) = \mu \tag{1.1}$$

e tem como uma das suas aplicabilidades em RNAs aproximadoras de universais de funções, visando mapear o comportamento entre as variáveis de entrada e saída de processos (SILVA; SPATTI; FLAUZINO, 2010). A função tangente hiperbólica assumirá sempre valores reais entre -1 e 1, tendo sua expressão matemática definida pela equação

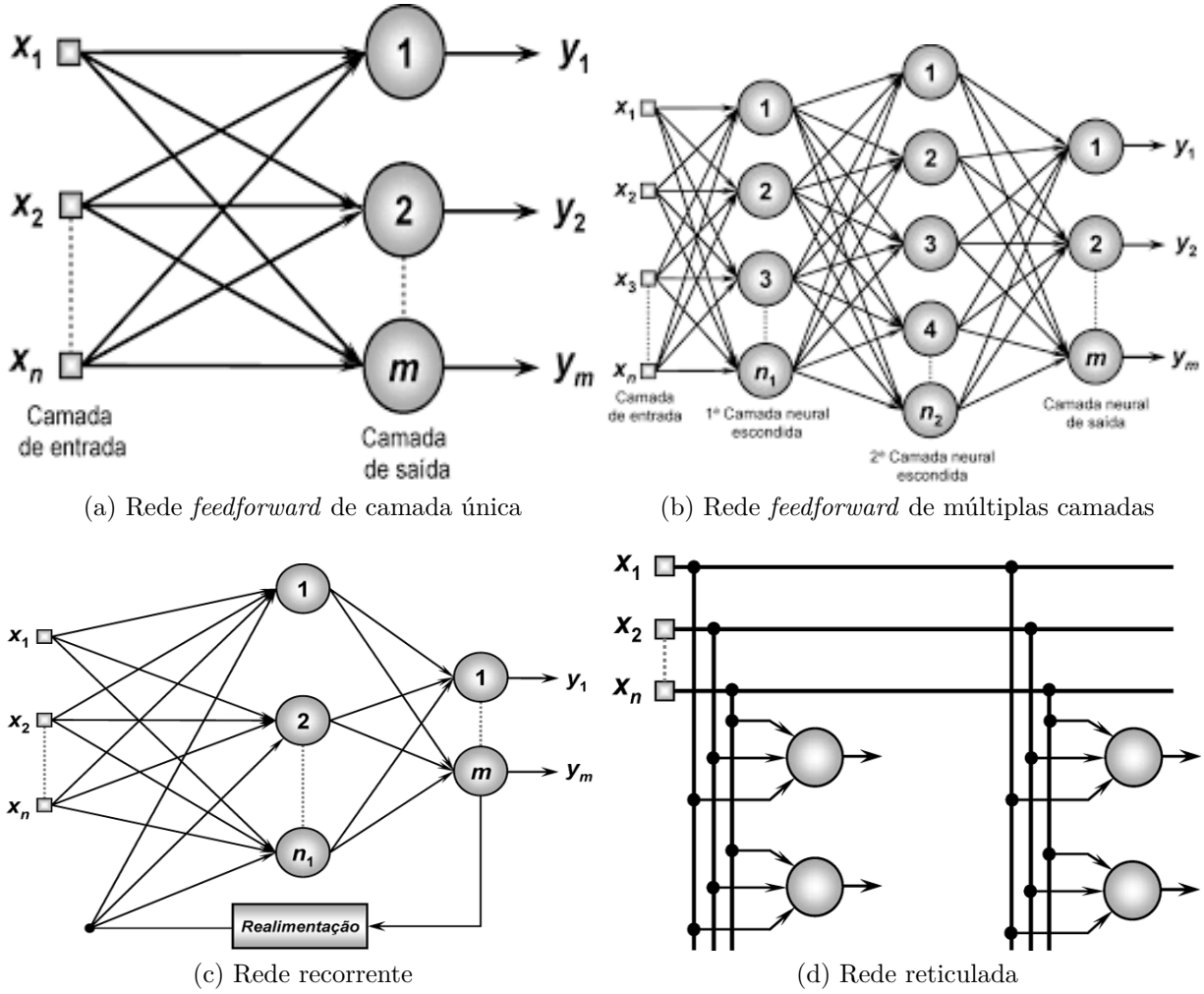
$$g(\mu) = \frac{1 - e^{-\beta\mu}}{1 + e^{-\beta\mu}} \tag{1.2}$$

onde β é uma constante real associada ao nível de inclinação da função (parâmetro de inclinação). Por fim, a função *softmax* (BRIDLE, 1990) normaliza entre 0 e 1 as saídas, com o objetivo de definir a probabilidade de uma classe dentro de um problema multi-classes, cuja sua expressão matemática é dada pela equação

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{c} e^{x_j}} \tag{1.3}$$

onde c é o número de camadas escondidas da RNA e x_i é o valor resultante do i-ésimo neurônio da camada escondida. O resultado da equação é interpretado como probabilidades, sendo úteis na classificação pois dá uma medida de certeza sobre as classificações do problema.

A arquitetura de uma RNA define a forma como os seus diversos neurônios estão arranjados ou dispostos, uns em relação aos outros. Elas podem ser de quatro tipos: *feed-forward* de camada simples, *feedforward* de múltiplas camadas, recorrentes ou reticuladas. Entende-se por rede *"feedforward"* àquela onde o fluxo de informações segue sempre em uma única direção, ou seja, da camada de entrada para a camada de saída. Uma RNA com arquitetura *feedforward* de camada simples tem apenas uma camada de entrada e uma única camada de neurônios, que é a camada de saída, mostrada na Figura 5a. Já uma RNA com arquitetura *feedforward* de múltiplas camadas contém pelo menos uma camada escondida de neurônios, como mostra a Figura 5b. A RNA com arquitetura recorrente tem suas saídas realimentadas como sinais de entrada para outros neurônios, mostrada na Figura 5c. Finalmente, uma RNA com arquitetura reticulada contém a localização espacial dos neurônios diretamente relacionada com o processo de ajuste de seus pesos e limiares, como mostra a Figura 5d.



(a) Rede *feedforward* de camada única

(b) Rede *feedforward* de múltiplas camadas

(c) Rede recorrente

(d) Rede reticulada

Figura 5 – Principais arquiteturas de RNAs (SILVA; SPATTI; FLAUZINO, 2010)

Já a topologia de uma RNA, considerando determinada arquitetura, pode ser definida como sendo as diferentes formas de composições estruturais que esta pode assumir. Um exemplo é que pode-se ter duas topologias pertencentes a uma mesma arquitetura, sendo que uma é composta por cinco neurônios e a outra é de 10 neurônios, ou pode-se alterar as suas funções de ativação, ou seu limiar de ativação, entre outros.

O treinamento de uma arquitetura específica consiste da aplicação de um conjunto de passos ordenados com o intuito de ajustar os pesos e os limiares de seus neurônios. Dessa forma, tal processo de ajuste (ou algoritmo de aprendizagem) visa sintonizar a rede para que as suas respostas estejam próximas dos valores desejados. Existem três tipos de treinamento: supervisionado, não-supervisionado e com reforço. O treinamento supervisionado visa em ter disponível as respectivas saídas desejadas, ou seja, cada amostra de treinamento é composta pelos sinais de entradas e suas correspondentes saídas. O treinamento não-supervisionado difere do supervisionado por não haver as respectivas saídas, forçando a rede a se auto-organizar em relação às particularidades existentes entre os elementos componentes do conjunto total de amostras. Já o treinamento com

reforço (SUTTON; BARTO, 1998) é uma variação do treinamento supervisionado e não-supervisionado, pois os parâmetros são ajustados baseando-se em quaisquer informações quantitativas ou qualitativas provenientes da interação com o sistema.

Entre os principais tipos de redes tendo arquitetura *feedforward* de camada simples estão o *Perceptron* (ROSENBLATT, 1958) e o *Adaline* (WIDROW; HOFF, 1988), cujos algoritmos de aprendizado utilizados em seus treinamentos são os baseados na regra de Hebb (HEBB, 1949) e na regra Delta (WIDROW; HOFF, 1988). Em relação às redes com arquitetura *feedforward* de múltiplas camadas, os principais tipos de redes são o *Perceptron* multicamadas (do inglês: *Multilayer Perceptron*, MLP) e as redes de base radial (do inglês: *Radial Basis Function*, RBF) (BROOMHEAD; LOWE, 1988). Já as principais redes com arquitetura recorrente são as redes de Hopfield (HOPFIELD, 1982) e a rede *Perceptron* com realimentação. Por fim, a rede de Kohonen (KOHONEN, 1982) é a principal representante das redes com arquitetura reticulada.

Assim, as características mais relevantes das Redes Neurais Artificiais, levando em conta todas as informações passadas neste capítulo, são:

- **Adaptação por experiência**, na qual os parâmetros internos (pesos sinápticos e limiar de ativação) são ajustados a partir da apresentação sucessiva de exemplos tais como padrões, amostras, medidas, os quais são relacionados ao comportamento do processo, possibilitando a aquisição do conhecimento por aquisição;

- **Capacidade de aprendizado** por meio de um processo de treinamento, cujo resultado é o relacionamento existente entre as diversas variáveis que compõem a aplicação;

- **Habilidade de generalização**, pois ao terminar o processo de treinamento da rede, essa é capaz de generalizar o conhecimento adquirido, sendo possível prever soluções até então desconhecidas;

- **Organização de dados**, onde a rede é capaz de realizar a sua organização interna, visando possibilitar o agrupamento de padrões que apresentam particularidades em comum, sendo baseada em características específicas envolvendo determinado conjunto de informações a respeito de um processo;

- **Tolerância a falhas**, uma vez que devido ao elevado nível de interconexões entre os neurônios artificiais, a rede neural torna-se um sistema tolerante a falhas quando parte de sua estrutura interna é sensivelmente corrompida;

- **Armazenamento distribuído**, uma vez que a busca do conhecimento a respeito do comportamento de determinado processo dentro de uma arquitetura neural é realizada de forma distribuída entre os diversos pesos sinápticos, permitindo um

incremento da robustez da arquitetura frente a eventuais neurônios que se tornaram inoperantes;

- **Facilidade de prototipagem**, uma vez que após o processo de treinamento da rede, os seus resultados são obtidos por operações matemáticas elementares (soma e multiplicação), permitindo a prototipagem das redes em *hardware* ou *software*.

As RNAs podem ser empregadas em diversos problemas relacionados às engenharias e ciências. Silva, Spatti e Flauzino (2010) citam algumas potenciais áreas de aplicabilidade, a saber:

- **Aproximador universal de funções**: o objetivo consiste em mapear o relacionamento funcional entre as variáveis (geralmente reais) de um sistema a partir de um conjunto conhecido de seus valores representativos. As aplicações para essa área são várias, sendo que envolvem normalmente o mapeamento de processos cuja modelagem por técnicas convencionais é de difícil obtenção;

- **Reconhecimento/classificação de padrões**: para esse tipo de aplicação, o objetivo consiste em associar um padrão de entrada, chamada amostra, para uma das classes previamente definidas, tal como ocorre em reconhecimento de imagens e voz, por exemplo. Assim, o problema a ser tratado possui um conjunto discreto e conhecido das possíveis saídas desejadas;

- **Agrupamento de dados (*clustering*)**: o objetivo é identificar e detectar similaridades e particularidades entre os diversos padrões de entrada para então possibilitar seu agrupamento;

- **Otimização de sistemas**: o alvo consiste em minimizar ou maximizar uma função custo (ou função objetivo) obedecendo também eventuais restrições que são impostas para o correto mapeamento do problema.

Vale ressaltar que neste trabalho foram utilizadas RNAs com arquitetura *feedforward* de múltiplas camadas e redes de base radial, pois, além de serem consideradas uma das arquiteturas mais versáteis quanto à aplicabilidade, o trabalho resolve problemas de reconhecimento e classificação de padrões, uma das potenciais áreas das RNAs MLP e RBF.

Informações mais detalhadas sobre Redes Neurais Artificiais podem ser encontradas em Parks, Levine e Long (1998), Kriesel (2007), Haykin (2009) e Samarasinghe (2016).

1.3 Ferramentas e Recursos

Nesta seção são apresentadas as ferramentas e recursos computacionais usados neste trabalho.

1.3.1 NLTK: *Natural Language ToolKit*

O NLTK, ou *Natural Language ToolKit* é uma biblioteca *open source*[2] presente na linguagem Python[3] que inclui extensivos dados, *software* e documentação gratuitamente disponível para *download*. Distribuições são fornecidas para Windows, Macintosh e plataformas Unix.

O livro de Bird, Klein e Loper (2012) foi utilizado como base de estudos para o entendimento da biblioteca NLTK. De acordo com os autores, o NLTK foi originalmente criado em 2001 como parte do curso de linguística computacional do Departamento de Computação e Ciência da Informação na Universidade da Pensilvânia (EUA). Desde então, tem sido desenvolvido e expandido com a ajuda de vários colaboradores, além de ter sido adotado em cursos de graduação e pós-graduação em dezenas de universidades.

A escolha da linguagem Python deu-se porque contém uma suave curva de aprendizado, devido a sua sintaxe e semântica serem transparentes e ter muitas funcionalidades para manipulação de *string*, o que é importante quando se trabalha com PLN utilizando textos. O NLTK define uma infraestrutura que pode ser usada para construir programas de PLN em Python, provendo classes básicas para representar dados relevantes, interfaces padrões para realizar tarefas como etiquetar textos (*Part-of-Speech Tagging*), análise sintática e classificação de textos. Implementações padrão de cada uma dessas tarefas podem ser combinadas para resolver problemas complexos.

Os principais módulos do NLTK e as funcionalidades mais utilizadas são mostradas na Tabela 1.

Tabela 1 – Principais módulos e funcionalidades do NLTK (BIRD; KLEIN; LOPER, 2012)

Tarefa de PLN	Módulo NLTK	Funcionalidade
Acessar Córpus	`nltk.corpus`	Padronizar as interfaces para Córpus e léxicos
Processamento de *strings*	`nltk.tokenize, nltk.stem`	Separar palavras por meio de tokenizadores, tokenizadores de sentenças e *stemmers*
Etiquetagem	`nltk.tag`	Definir as classes gramaticais das palavras

[2] Código aberto: distribuição livre, código fonte, trabalhos derivados, distribuição da licença entre outros
[3] http://www.python.org

1.3.1.1 Etiquetador *Mac-Morpho*

Um dos módulos presentes no NLTK é a etiquetagem (*Part-of-Speech Tagger* ou *POS Tagger*). Segundo Bird, Klein e Loper (2012), um etiquetador é responsável pelo processo de definição da classe gramatical das palavras, de acordo com as funções sintáticas. As principais classes no Português são os substantivos, pronomes, adjetivos, verbos, advérbios, preposições, conjunções, numerais e interjeições. Geralmente o resultado do processamento de um etiquetador é dado por conjunto de *tags*.

No NLTK, o etiquetador usado para processar palavras em Português é o *Mac-Morpho*, proposto por Aluísio et al. (2003). Eles apresentam um Córpus de 1.1 milhões de palavras validadas manualmente com anotação morfossintática. O conjunto de texto base foi retirado do jornal Folha de São Paulo, onde de acordo com os autores, retorna uma alta qualidade contemporânea pelo fato de englobar diferentes autores e domínios.

Sua documentação está disponível para livre acesso no site do NILC/USP[4]. Dentre o conjunto de etiquetas presentes no *Mac-Morpho*, algumas variações sintáticas estão explicitadas para melhores resultados, como a possibilidade de identificar tipos diferentes de pronomes (substantivos, pessoais, conectivos), advérbios (conectivos, relativos), conjunções (coordenativa, subordinativa), tipos de verbos, além de extras como estrangeirismos, apostos, datas, entre outros. As etiquetas usadas pelo *Mac-Morpho* são descritas no Anexo A.

Como exemplo, quando a frase "O hotel é bom" é submetida para análise, primeiramente o NLTK separa as palavras em *tokens* para depois dar uma etiqueta para a palavra. No caso da frase descrita, o conjunto de *tags* resultante do Mac-Morpho seria [('O', 'ART'), ('hotel', 'N'), ('é', 'V'), ('bom', 'ADJ')]. A lista de *tags* se torna manipulável para executar as técnicas para extrair as características e palavras opinativas.

1.3.2 Athena

A Athena[5] é uma ferramenta proposa por Oliveira et al. (2014) que oferece uma maneira simples para desenvolver sistemas de Inteligência Computacional (IC), utilizando um editor gráfico que propicia a programação visual das técnicas de IC. A partir disso, os autores criaram um novo conceito, chamado de *Computational Intelligence as a Service* (CIaaS).

Adicionalmente, a utilização da Athena permite o compartilhamento de sistemas por outros pesquisadores para que sejam possíveis discussões sobre algoritmos selecionados e até mesmo sobre os resultados obtidos.

A Figura 6 mostra o editor gráfico disponível na ferramenta. No lado esquerdo

[4] http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf
[5] Disponível em http://athenaufpi.appspot.com/

se encontram os módulos de IC disponíveis (*Components*). Os módulos são detalhados em Oliveira (2016) e são categorizados de acordo com a sua função ou subárea da IC ao qual pertencem (35 módulos estão disponíveis na Athena). Ao clicar em um módulo, ele será incluído na área de edição, localizada no centro da tela do editor. No lado direito do editor se encontra o menu de contexto, onde o usuário pode incluir ou excluir entradas e saídas dos módulos presentes na área de edição (*Configuration*). Existe a barra de ações presente no canto superior direito da tela, que proporciona ao usuário manipular a arquitetura ao salvá-la (*Save*) executá-la (*Run*), além de limpar a tela (*Reset*), voltar (*Back*) e exibir ajuda (*Help*).

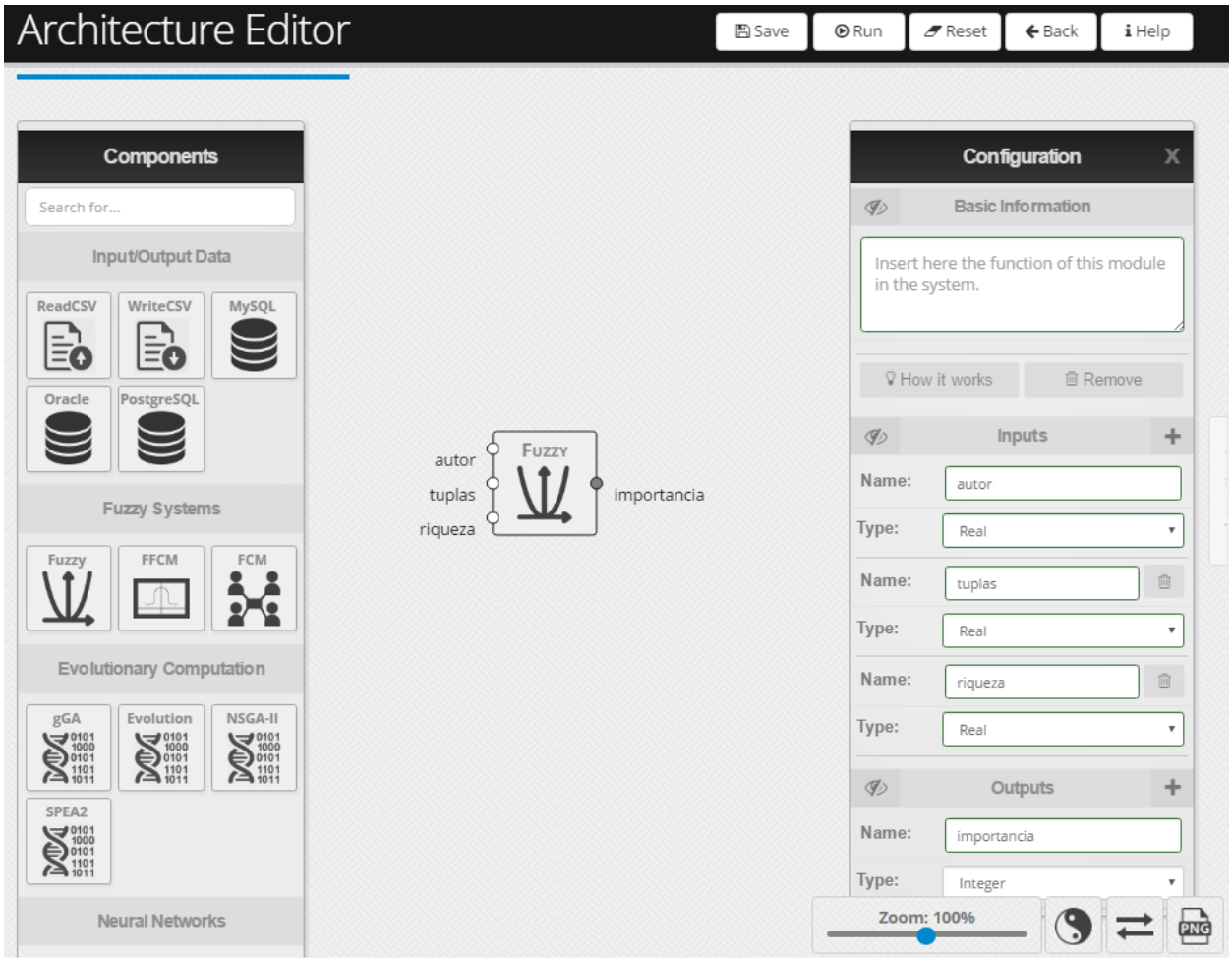


Figura 6 – Editor gráfico da Athena

A Athena foi escolhida para executar os sistemas pelo fato de oferecer suporte à resolução de Sistemas de Inferência *Fuzzy* e Redes Neurais Artificiais com facilidade. Adicionalmente, a Athena é uma arquitetura baseada em módulos que torna possível encapsular a complexidade interna das técnicas, deixando a mostra apenas o que interessa ao usuário final, a saber: entradas, configurações e saídas.

1.3.3 Córpus *Booking.com*

Por mais que existam vários trabalhos na área de mineração de opinião na língua portuguesa, ainda há ausência de recursos para a utilização pelos pesquisadores de PLN, diferente da língua inglesa, onde existem recursos disponíveis e de fácil acesso pelos pesquisadores. Em consequência dessa falta de recursos, especialmente quando se refere a um Córpus de comentários de consumidores sobre tal domínio, foi necessário coletar e construir um Córpus anotado com informações de polaridade dos comentários.

Para este trabalho, os comentários foram referentes às procuras por hotéis no Rio de Janeiro visando os Jogos Olímpicos Rio 2016. Dessa forma, foram coletados 9.909 comentários de cinco hotéis do site *Booking.com*, escolhidos aleatoriamente, no período de setembro de 2015 a agosto de 2016. Neste trabalho, a ferramenta de coleta automática utilizada foi o *Scrapy*[6] como *Web crawler*.

A estrutura de comentários do *Booking.com* é composta das seguintes informações, mostradas na Figura 7.

1. Autor da opinião;
2. Quantidade de avaliações que o autor fez no site *Booking.com*;
3. Nota dada para o hotel;
4. Comentário geral sobre a sua estadia;
5. Data do comentário;
6. Comentários negativos;
7. Comentários positivos;

Para a avaliação das abordagens propostas nesta Dissertação, foi necessário formar um Córpus com a avaliação da importância dos comentários. O procedimento de análise manual foi realizado por cinco especialistas, todos alunos de mestrado do curso de Letras Português da UFPI, utilizando o mesmo método de avaliação do Córpus revisado. Os especialistas avaliaram manualmente cada um dos comentários atribuindo um grau de importância, sendo: Insuficiente (ISF), Suficiente (SF), Bom (BM) ou Excelente (EXC), de acordo com a constatação de características sobre o hotel, riqueza do comentário e observação da reputação do autor pelo número de comentários realizados no site *Booking.com*.

Foi retirado um subconjunto de 370 comentários do Córpus revisado para a avaliação manual. O número de comentários foi escolhido por meio de uma análise estatística que leva em consideração o nível de confiança e a margem de erro da amostra. Sendo q a

[6] Disponível em http://scrapy.org/

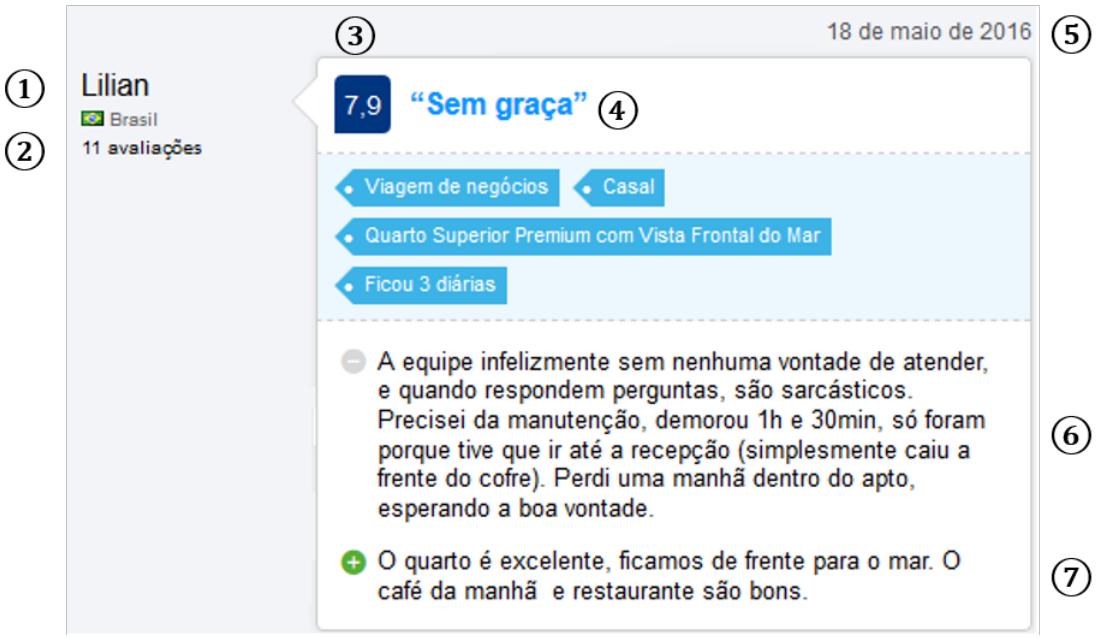


Figura 7 – Comentário completo do *Booking.com*

quantidade mínima de comentários pertencentes à amostra, o cálculo é realizado por meio das equações (HAMBURG, 1985)

$$x = Z\left(\frac{c}{100}\right)^2 r(100 - r) \tag{1.4}$$

$$q = \left[\frac{Tx}{((T-1)E^2 + x)}\right]^2 \tag{1.5}$$

onde $Z\left(\frac{c}{100}\right)$ é o valor crítico do nível de confiança c escolhido, r é a distribuição de respostas, ou seja, quais respostas são esperadas para cada comentário — geralmente usa-se 0,5, uma vez que se a amostra for altamente distorcida, a população provavelmente será, T é o tamanho do Córpus e E é a margem de erro. Para esse trabalho o nível de confiança escolhido foi de 95% e a margem de erro escolhida foi de 5%.

O resultado da análise manual é mostrado na Tabela 2. Para fins de definição, esse subconjunto com 370 comentários analisados foi nomeado como **Subcórpus de Referência de Importância**.

Tabela 2 – Distribuição das importâncias após a análise manual

Grau de Importância	Positivos	Negativos	Neutros	Total
Excelente	8	4	8	**20**
Bom	41	31	26	**98**
Suficiente	36	50	37	**123**
Insuficiente	36	59	34	**129**

Os comentários coletados continham dados interessantes de serem descritos. Dentre os dados descritivos dos comentários, destacam-se que:

- a nota média dada pelos autores foi de 7,91 (máximo 10,0). A média de nota geral dos cinco hotéis foi de 7,88 (maior nota foi do hotel 3 com 8,1 e a menor nota foi do hotel 4 com 7,5);
- o número de avaliações médio dos autores no site *Booking.com* foi de 4,1 e 1468 autores se identificaram como "Anônimo";
- 1757 comentários (17,73%) não tinham o campo referente ao comentário **negativo** preenchido. A nota média dada com essa característica foi de 8,92;
- 788 comentários (7,95%) não tinham o campo referente ao comentário **positivo** preenchido. A nota média dada com essa característica foi de 7,06;
- na base ainda tinham 3 comentários (0,03%) que não tinham nenhum dos dois campos preenchidos. A nota média dada com essa característica foi de 9,16;
- os comentários tiveram média de 33,4 palavras, sendo que o maior comentário continha 395 palavras. 99 comentários (0,99%) que continham apenas duas palavras, sendo os menores comentários da base. Nestes comentários não havia o campo referente à descrição **negativa** (82 comentários) ou **positiva** (17 comentários) preenchidas;
- os comentários gerais mais utilizados foram:
  - **Bom** com 541 comentários (5,45%), além de mais 293 comentários (2,95%) que continham o advérbio *muito*;
  - **Excepcional** com 296 comentários (2,98%). A nota média dada pelo autor com essa característica foi de 9,78, a maior entre todas pesquisadas com pelo menos 50 registros;
  - **Localização** com 212 comentários (2,13%).

A Figura 8 mostra o gráfico da distribuição dos comentários por sua data de postagem no site *Booking.com*. Percebe-se no gráfico que dois picos de comentários são bem visíveis: em Janeiro de 2016 e em Agosto de 2016. Ambos os picos são referentes aos hóspedes que escreveram o comentário após a estadia, o que explica o pico de Janeiro 2016, ligado ao *Réveillon*, uma vez que o Rio de Janeiro é um dos principais destinos dos turistas com lotação de 93% no período[7]. O pico de Agosto de 2016 é referente aos Jogos Olímpicos Rio 2016.

[7] http://www.brasil.gov.br/turismo/2015/12/brasil-e-destino-preferido-para-festa-do-ano-novo

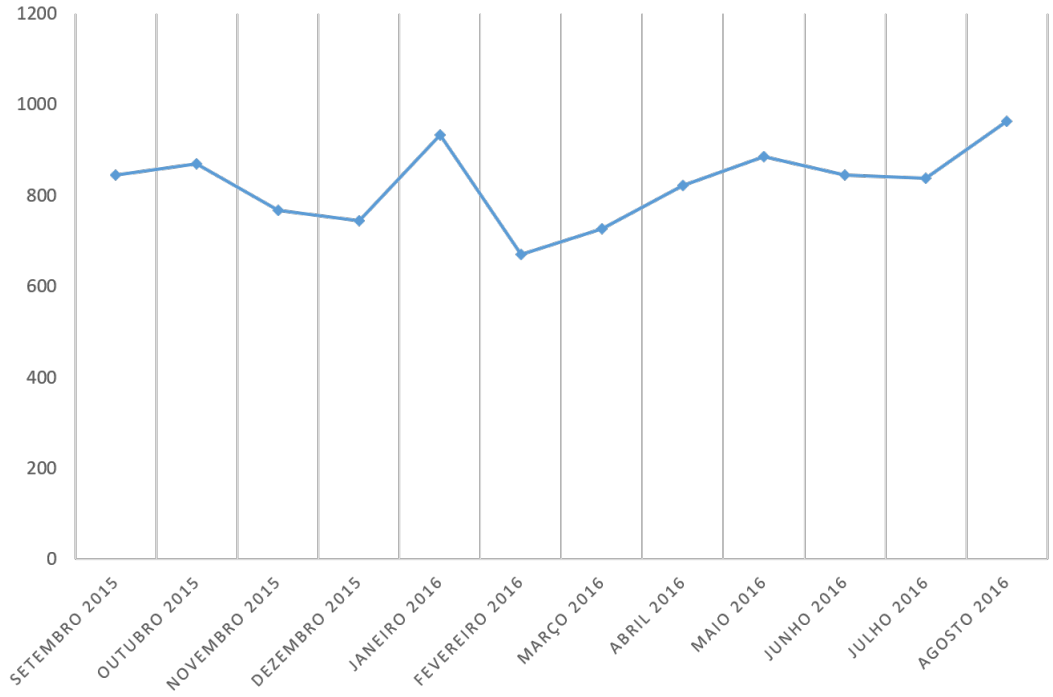


Figura 8 – Data de postagem dos comentários no site *Booking.com*

Por fim, vale ressaltar que o *Booking.com* separa os comentários em positivos e negativos. O autor tem a possibilidade de preencher nos campos relacionados os comentários positivos e negativos, mas nem todos o fazem. Como mostrado nos dados descritivos acima, aproximadamente 25% dos autores não preencheram um dos campos. O arquivo `.sql` contendo a tabela referente ao Córpus *Booking.com* usado neste trabalho está livremente disponível para utilização[8].

1.4 Considerações Finais

Este capítulo apresentou o referencial teórico baseado nos principais conceitos que envolvem a pesquisa discutida nesta Dissertação, tais como Mineração de Opinião e considerações sobre Sistemas *Fuzzy* e Redes Neurais Artificiais. Além desses assuntos, foram discutidas as ferramentas e recursos utilizados na implementação das abordagens propostas e nos experimentos realizados.

[8] Disponível em https://goo.gl/5ZGl4M

2 Trabalhos Relacionados

Como explicado na seção 1.1.1, a mineração de opinião consiste em três etapas: i) identificar a opinião; ii) analisar e classificar a opinião; e iii) sumarizar os resultados. Existem trabalhos que tratam apenas de uma das partes separadamente e trabalhos que englobam todas as partes em um único módulo. Este trabalho contempla a etapa de identificação, mas podendo ser abrangido para as outras etapas. A seguir serão apresentados os trabalhos referentes a cada uma das etapas citadas.

2.1 Identificação

Para identificar a opinião, os trabalhos exploram: i) tópicos; ii) conteúdo subjetivo; e iii) entidades ou aspectos. As principais dificuldades encontradas são a co-referência, resolução de pronomes, tratamento de negação e tratamento de ironia/sarcasmo. As principais técnicas utilizadas para tentar resolver as dificuldades citadas são o reconhecimento de padrões sintáticos e n-gramas, eliminando os termos irrelevantes. As ontologias podem ser usadas para definir características importantes e que devem ser consideradas para a análise e classificação.

Em um dos principais trabalhos da área, Turney (2002) define cinco padrões linguísticos utilizados na extração de frases contendo adjetivos e verbos. Essas classes gramaticais demostram ser bons indicadores de subjetividade e sentenças avaliativas para definir a classificação do sentimento (HATZIVASSILOGLOU; WIEBE, 2000). Os padrões linguísticos, para o inglês, são mostrados na Tabela 3. Uma adaptação desses padrões foi feita para serem utilizados neste trabalho, a qual será explicada em detalhes na seção 3.2.1.

Tabela 3 – Padrões de Turney (TURNEY, 2002)

Padrão	1º palavra	2ª palavra	3ª palavra
1	ADJ	SUBST	qualquer classe
2	ADV	ADJ	não SUBST
3	ADJ	ADJ	não SUBST
4	SUBST	ADJ	não SUBST
5	ADV	VERBO	qualquer classe

O trabalho de Liu, Wu e Yao (2006) propõe as três partes da mineração de opinião e explora o quesito de entidade e aspectos. Primeiro eles identificam todas as expressões relacionadas com o domínio para classificar em dois grupos: características e produtos. Os autores apresentam um algoritmo para prever a dependência entre características e produtos, onde todas as opiniões são indexadas como uma tripla <produto, característica,

qualidade> e em seguida as tuplas são utilizadas para recuperar opiniões que "casam" com os interesses dos usuários.

Ding, Liu e Yu (2008) propuseram um método efetivo para identificação das orientações semânticas de opiniões em inglês expressadas por consumidores sobre características de produtos. O método é capaz de lidar com dois principais problemas dos métodos existentes: i) palavras opinativas cujas operações semânticas são dependentes do contexto, por exemplo: no domínio de câmeras, em *"the battery life is very long"* (tradução livre: o tempo de vida da bateria é muito longo) e *"it takes a long time to focus"* (tradução livre: leva um tempo longo para focar), a palavra opinativa *"long"* pode significar uma opinião positiva ou negativa, respectivamente; e ii) agregar múltiplas palavras opinativas em uma mesma sentença. Para resolver o problema i), eles propuseram uma abordagem holística[1] que pode precisamente inferir a orientação semântica de uma palavra opinativa baseada no contexto da opinião ou de outra opinião que aparece a mesma palavra opinativa. Para resolver o problema ii), eles usaram uma nova função para combinar múltiplas palavras opinativas em uma mesma sentença. Eles também consideraram tanto opiniões explícitas quando implícitas, onde o método deles também lida com características implícitas representadas por indicadores das características.

Jeong, Shin e Choi (2011) propõem um sistema para extração e refinamento de características baseando-se em sintagmas nominais e informações semânticas das palavras, o que denominaram de FEROM. Devido a problemas que levavam a resultados insatisfatórios na extração de aspectos na época, os autores propuseram uma fase de pré-processamento onde todas as palavras dos comentários são etiquetados com suas classes gramaticais para então identificar os sintagmas nominais. Além disso, na mesma etapa de pré-processamento há um separador de sentenças, que mantém cada aspecto e suas palavras opinativas em uma mesma frase.

Já Silva, Lima e Barros (2012) apresentam o *SAPair*, um processo de análise de sentimento mais refinado, entrando no nível de características. A proposta foi classificar a polaridade das opiniões sobre cada característica do objeto sendo monitorado, por meio dos pares <característica, palavra opinativa> , uma vez que alguns adjetivos mudam de polaridade quando acompanham um certo substantivo, por exemplo, "pizza quente" e "cerveja quente", de cunho positivo e negativo, respectivamente. Experimentos mostraram que o processo proposto tem alta eficácia, no qual supera outros métodos existentes, tal como a proposta de Turney (2002).

Shi e Li (2011) primeiramente analisaram pesquisas anteriores referentes à classificação de sentimentos sobre comentários de hotéis e então propuseram uma abordagem com aprendizado de máquina supervisado usando unigramas com dois tipos de informação: a

[1] Privilegia o todo ou um sistema, e não as suas partes componentes tomadas isoladamente (XIMENES, 2000)

frequência e a medida TF-IDF (LUHN, 1957; JONES, 1988). As informações eram usadas para realizar a classificação de polaridade dos documentos. Com o valor de medida-F 87,2%, o TF-IDF foi mais efetivo.

O trabalho de Aciar et al. (2007) faz uso de ontologia para propor um sistema de recomendação de produtos. A ontologia tem a função de mapear a qualidade de opiniões e de produtos em um formato próprio para o procedimento de recomendação. Eles também utilizam o tratamento de características sinônimas e implícitas, onde cada característica da ontologia possui uma lista de palavras relacionadas, identificando o aspecto referente.

2.2 Classificação

Com relação à classificação, normalmente explora-se a classificação binária: positivo ou negativo. Classes adicionais podem ser consideradas para aumentar o nível de detalhes da análise, por exemplo: i) a classificação por estrelas, que pode ir de 0 a 5; ii) por nota, que pode ir de 0 a 10; iii) valores diversos, como separação do comentário como excelente, bom, regular e ruim. Quanto às dificuldades de classificação, se destacam: uso de palavras de sentimento pode ser enganoso, uso de ironia/sarcasmo em muitos domínios, opinião pode depender do observador e polaridade do comentário nem sempre é objeto de consenso, por exemplo, o consenso entre avaliadores humanos gira em torno de 75% (BRUCE; WIEBE, 1999; PANG; LEE; VAITHYANATHAN, 2002).

As principais técnicas de classificação são: i) uso de léxico (dicionário) com palavras ou expressões de sentimentos, por exemplo, Sentilex-PT (SILVA; CARVALHO; SARMENTO, 2012) e *OpLexicon* (SOUZA et al., 2011); e ii) aprendizagem de máquina, por exemplo, Máquinas de Vetores de Suporte (do inglês *Support Vector Machine*, SVM) (MULLEN; COLLIER, 2004; TAN et al., 2011), Sistemas *Fuzzy* (DRAGONI; TETTAMANZI; PEREIRA, 2014; INDHUJA; REGHU, 2014; SOUSA, 2015) e Redes Neurais Artificiais (SHARMA; DEY, 2012a; SHARMA; DEY, 2012b).

Dentre os léxicos existentes na literatura, o *WordNet* (FELLBAUM, 1998) é o maior e o mais conhecido. Alguns métodos existentes o utilizam como base para criar outros léxicos mais específicos. Existe uma adaptação da versão em inglês para o Português brasileiro chamada *WordNet.BR* (DIAS-DA-SILVA, 2010). Léxicos de sentimento são usados para definir a classificação semântica do comentário. Um importante léxico de sentimentos para o Português é o Sentilex-PT (SILVA; CARVALHO; SARMENTO, 2012), contendo 7014 *lemmas* e 82.347 formas flexionadas.

Em relação à orientação semântica por meio de léxico de sentimentos, Kamps et al. (2004) utilizaram as relações semânticas presentes na *WordNet*. Um grafo com os adjetivos contidos na intersecção entre uma lista de termos e a *WordNet* foi definido, adicionando um link entre dois adjetivos sempre que a *WordNet* indicar uma relação de sinônimos

entre eles. Além disso, os autores definiram as palavras "*good*" e "*bad*" como palavras de referência quando um termo é apresentado. A orientação semântica desse termo é calculada de acordo com a sua distância relativa em comparação com as duas palavras de referência. O termo é considerado positivo se a distância relativa for um valor positivo e vice-versa. Esse método tornou possível definir a orientação semântica de sinônimos dos adjetivos.

No âmbito da aprendizagem de máquina, mais precisamente utilizando Sistemas *Fuzzy*, ainda existem poucos trabalhos em métodos de mineração de opinião. Fu e Wang (2010) apresentam um *framework* utilizando conjuntos *Fuzzy* para classificar sentimentos a nível de sentença. Uma função de pertinência é usada para identificar o grau de pertinência de cada sentença, classificando os conjuntos: Positivo, Negativo e Neutro, representando as classes de polaridades de sentimentos. Kar e Mandal (2011) propuseram um sistema de mineração de opiniões que visa definir a polaridade e intensidade do sentimento de opiniões por meio do uso de conjuntos *Fuzzy*. Nadali, Murad e Kadir (2010) criaram um sistema *Fuzzy* que executa a classificação de sentimentos de opiniões de consumidores em variações de positivo e negativo (muito positivo e negativo, moderadamente positivo e negativo e muito fracamente positivo ou negativo).

A abordagem de Sousa (2015) utilizou a estrutura sintática das sentenças para identificar as características e seus respectivos qualificadores por ser simples e de fácil implementação. O autor considerou os verbos como palavras opinativas, além dos adjetivos e advérbios e utilizou padrões linguísticos pré-definidos por Turney (2002) e algumas extensões para satisfazer o domínio de produtos nos quais ele trabalhou. O método para inferência da orientação semântica das expressões faz o uso do léxico de sentimentos Sentilex-PT. Na abordagem, foi utilizado Sistema *Fuzzy* para inferir a importância dos comentários avaliados, apresentando como principal vantagem permitir classificá-los e definir os comentários mais significantes, reduzindo a complexidade da tarefa de avaliar os inúmeros comentários de produtos e serviços. No Sistema *Fuzzy*, foram usados como entrada: a reputação do autor, a quantidade de padrões linguísticos e o percentual de palavras corretamente escritas nos comentários, definida como riqueza do vocabulário.

Santos et al. (2016a) realizaram um estudo experimental que envolvia a abordagem de Sousa (2015) utilizando Sistemas *Fuzzy* com uma abordagem similar, porém utilizando uma Rede Neural Artificial. Os mesmos experimentos definidos em Sousa (2015) foram realizados a fim de obter uma comparação justa entre as abordagens. Uma RNA MLP com o algoritmo de aprendizagem *backpropagation* (RUMELHART; HINTON; WILLIAMS, 1986) foi utilizada com sua topologia definida através da validação cruzada. A mesma amostra de comentários foi utilizada no funcionamento da RNA, gerando resultados inferiores ao da abordagem base. Os resultados desse trabalho sugeriram outros experimentos, resultando nesta Dissertação.

2.3 Sumarização

Por fim, a etapa de sumarização consiste em criar métricas e sumários que quantificam a diversidade das opiniões a respeito de um alvo em cada um dos níveis de análise textual, tais como documento, sentença e entidade ou aspecto específico.

Wang e Liu (2015) exploraram a sumarização de opinião sobre conversas espontâneas usando dois tipos de abordagens: uma supervisionada e outra não-supervisionada, sendo esta última baseada em métodos gráficos. Além disso, os autores investigaram o uso da resolução de pronome na tarefa de sumarização, desenvolvendo várias características baseadas na pronúncia de co-referência e incorporando-as em seu sistema de sumarização de opinião supervisionada. Resultados experimentais mostraram que ambas as abordagens superaram a abordagem base e os recursos relacionados ao pronome podem ajudar a gerar melhores sumários.

Ainda a nível de entidade ou aspecto, Kurian e Asokan (2015) propuseram um método para a sumarização de opinião em domínios de produtos para os quais os dados pré-etiquetados não estão disponíveis. O método proposto utiliza a classificação de sentimentos entre domínios (do inglês: *Cross-Domain Sentiment Classification*) para a criação de sumarizações gráficas sensíveis ao aspecto. A classificação de sentimentos entre domínios utiliza as informações de sentimento obtidas de outro domínio de produto para prever a classe de opinião de opinião no domínio de destino, considerando o contexto em que a palavra opinativa é usada. O método de classificação de sentimentos entre domínios alcançou exatidão quando comparada com a abordagem baseada na *SentiWordNet* (ESULI; SEBASTIANI, 2006). Experiências conduzidas com diferentes domínios mostraram que quando há a consideração de domínios de origem semelhantes aos domínios, o desempenho do método melhora.

Com relação ao nível de sentença, Kolekar et al. (2016) mostraram um sistema de análise de sentimentos e classificação utilizando PLN, técnicas de aprendizagem de máquina e uma abordagem baseada em dicionário onde a metodologia proposta classifica o sentimento nas diferentes classes de polaridade (positivo, negativo e neutro) apresentando uma solução viável para o modelo *bag-of-words* (BOW), onde os autores conseguiram eliminar e modificar o deslocador de polaridade de negação de um determinado texto. Por exemplo: na frase *"I don't like this digital camera. Picture quality of this camera is excellent. It is too expensive"* (tradução livre: "Eu não gostei desta câmera digital. Qualidade da imagem desta câmera é excelente. É muito caro"), os autores detectam o deslocador de polaridade de negação com sendo *"I don't like this digital camera"* e quando ocorre a eliminação e modificação desse deslocador, a sentença fica *"I dislike this digital camera"*. A expressão *don't* é detectada como uma mudança na orientação semântica, assim sendo substituída por um antônimo referente à palavra opinativa *dislike*.

Quando se trata de sumarização de opiniões referentes a hotéis, alguns trabalhos estão presentes na literatura. Raut e Londhe (2014a) usaram o método de pontuação de frequência e relevância de termos para representar as frases mais informativas na sumarização. Já Wang, Lu e Zhai (2010) apresentaram a Análise de Classificação de Aspecto Latente (do inglês: *Latent Aspect Rating Analysis*, LARA), que visa analisar opiniões expressas sobre uma entidade em um comentário no nível de aspectos, permitindo a sumarização da opinião em sua polaridade e o *ranking* de entidades com base na classificação dos aspectos.

Um bom trabalho que trata da comparação de métodos para sumarizações em geral é apresentado por McNeill et al. (2015) e uma comparação de métodos de sumarização de opiniões é tratado por Condori e Pardo (2017). Outras pesquisas (*surveys*) que contém mineração de opinião e métricas para sumarização da informação são encontradas em Khan et al. (2009), Seerat e Azam (2012), Vinodhini e Chandrasekaran (2012) e Raut e Londhe (2014b).

2.4 Considerações Finais

Vale ressaltar que o foco desta Dissertação está na identificação e extração de características, tal como proposto por Silva, Lima e Barros (2012), porém, para o Português brasileiro. Consequentemente, as outras etapas da mineração de opinião (classificação e sumarização) não são plenamente exploradas. Além disso, o que se propõe nesta Dissertação difere dos trabalhos apresentados por utilizar uma combinação de padrões linguísticos e uma base de características na identificação das características e outros índices de riqueza léxica na avaliação da riqueza do vocabulário. O diferencial em relação à abordagem TOP(X) de Sousa (2015) está nas adaptações nas variáveis referentes à quantidade de tuplas e riqueza do vocabulário. Destaca-se que a variável referente à reputação do autor está sendo explorada por outro pesquisador do grupo de PLN da UFPI.

3 Abordagem com Sistema *Fuzzy* e Rede Neural Artificial

Considerando que o objetivo principal deste trabalho é fazer adaptações em duas variáveis de entrada da abordagem TOP(X) proposta por Sousa (2015), além de propor uma comparação entre abordagens que utilizam Sistema *Fuzzy* com abordagens que utilizam RNA, neste Capítulo são apresentadas as abordagens propostas. É importante destacar que uma nova versão da abordagem TOP(X) foi implementada, para que os etiquetadores, *stemmers* e dicionários fossem os mesmos nas abordagens de referência, evitando, assim, erros decorrentes do uso de diferentes ferramentas e recursos de PLN. Primeiramente é destacada a abordagem TOP(X), em seguida são destacadas as adaptações nas variáveis de entrada, quantidade de tuplas e riqueza do vocabulário, implementadas nesta versão e por fim são discutidas as abordagens propostas nesta Dissertação.

3.1 Abordagem TOP(X) Original

Sousa (2015) propôs uma abordagem para estimar o grau de importância de comentários sobre produtos e serviços de usuários da Web. Para alcançar esse objetivo, ele propôs um sistema de inferência *Fuzzy* com três entradas: reputação do autor, número de tuplas <característica, palavra opinativa> e corretude do vocabulário, denominado analisador de riqueza e a saída do sistema é o grau de importância do comentário (ver Figura 1). Essa abordagem será denominada nesta Dissertação de **abordagem base**.

Para uma comparação justa com as abordagens propostas nesta Dissertação, foi necessário implementar uma nova versão da abordagem TOP(X) que utilizou um outro etiquetador, outros padrões linguísticos e outras funções de pertinência e base de regras, a serem discutidas nas adaptações e abordagens durante este capítulo. A ideia da abordagem continua a mesma, mas era imprescindível alterar os recursos para que a comparação fosse feita sem erros decorrentes das ferramentas utilizadas.

3.2 Adaptações

As variáveis utilizadas por Sousa (2015) para compor a abordagem TOP(X) foram: reputação do autor, quantidade de tuplas e riqueza do vocabulário. O autor destaca que essas variáveis foram escolhidas de forma empírica e ele acredita que sejam as principais variáveis para definir a importância de um comentário postado na *Web*, pois um comentário informativo deve ser escrito em Português correto, com grande cobertura de características

do produto ou serviço e realizado por um autor confiável.

A variável reputação do autor é calculada pelo número de comentários que o autor realizou em todo o *site*. Sabe-se que a maneira de calcular a reputação do autor não é adequada pois um usuário pode postar vários comentários sem utilidade com o objetivo de aumentar esse número e, assim, se tornar um autor com boa reputação.

A variável quantidade de tuplas é calculada a partir do número de tuplas <característica, palavra opinativa> encontradas no texto. Uma rotina para a extração de tuplas é executada utilizando reconhecimento de padrões linguísticos que resulta na quantidade de características e palavras opinativas presentes no comentário. Por fim, a variável riqueza do vocabulário é definida pelo percentual de palavras escritas corretamente no comentário.

Nesta Dissertação, a variável reputação do autor é a quantidade de comentários que o usuário fez no site *Booking.com*. Porém, é importante mencionar que um pesquisador do grupo de pesquisa está explorando novas medidas para definir a reputação do autor. As adaptações para melhorias ocorreram somente nas variáveis referentes à quantidade de tuplas e à riqueza do vocabulário e são discutidas nas subseções a seguir.

3.2.1 Adaptações na variável quantidade de tuplas

Sousa (2015) extrai as tuplas <característica, palavra opinativa> verificando a estrutura frasal "sujeito + verbo de ligação + predicativo" no qual o núcleo do sujeito define a característica e a núcleo do predicativo indica a palavra de qualidade. Além disso, ele utiliza padrões linguísticos pré-definidos baseados nos padrões de Turney (2002) para a extração das tuplas por meio do reconhecimento dos padrões linguísticos no domínio de *smartphones*.

As adaptações realizadas nesta variável foram necessárias, uma vez que o domínio de aplicação foi diferente. Assim, uma análise manual em um conjunto de comentários foi realizada por três especialistas da área de linguística, os quais dois deles são alunos de graduação e um aluno de mestrado do curso de Letras-Português da UFPI. Por meio da análise manual foi possível identificar características[1], palavras opinativas, advérbios e verbos importantes presentes nos comentários. Dessa forma, permitiu-se construir uma base contendo as principais características de hotéis.

Proveniente da análise manual foram encontradas 51 características. Adicionalmente, o site *Booking.com* define 7 características em cada página de hotel. Com o intuito de aumentar o escopo da base de características, foi utilizado o *ConceptNet* (LIU; SINGH, 2004) para coletar os sinônimos das características previamente encontradas. Nele foram encontradas mais 28 características. Uma filtragem manual foi realizada a fim de eliminar

[1] Neste trabalho considera-se como característica as entidades, aspectos e seus atributos.

as características que continham o mesmo *stemming*[2], finalizando a base de características com 81 palavras. Esta base de características é mostrada no Apêndice A está livremente disponível para utilização[3].

A construção da base de características permitiu que no processo de etiquetagem das palavras dos comentários, as características detectadas fossem apenas aquelas que fazem parte da base de características criada. Este método permitiu maior controle das características relacionadas a hotel e uma melhor acurácia nas tuplas que realmente descrevem aspectos de hotel.

A análise manual também contribuiu para a criação dos padrões linguísticos utilizados nesta Dissertação. Ainda com base nos padrões de Turney (2002), novos padrões linguísticos foram propostos para a extração das tuplas da abordagem. A Tabela 4 mostra os padrões propostos.

Tabela 4 – Padrões linguísticos identificados

1	(<SUBS> | <SUBS> <PREP> <SUBS>) <ADV>? <V>? <ADV>? <ADJ>
2	(<SUBS> | <SUBS> <PREP> <SUBS>) <V>? (<ADV>? <ADJ> <,>)* <ADV>? <ADJ> <CONJ> <ADV>? <ADJ>
3	<ADJ> (<SUBS> | <SUBS> <PREP> <SUBS>)
4	<ADV> <V> <PREP> (<SUBS> | <SUBS> <PREP> <SUBS>)
5	<V> <ADV> <PREP> (<SUBS> | <SUBS> <PREP> <SUBS>)

O padrão 1 coleta as características e palavras opinativas presentes no comentário, contendo ou não advérbios antes do adjetivo. Caso exista um verbo que ligue a característica com a palavra opinativa, o padrão 1 também reconhece a estrutura "sujeito + verbo de ligação + predicativo". "A localização é maravilhosa" e "atendimento muito rápido" são exemplos de detecção das tuplas pelo padrão 1.

O padrão 2 coleta as características com uma ou mais palavras opinativas relacionadas. O tratamento desse padrão se dá pelo número de adjetivos encontrados próximos à característica, levando em conta também os advérbios que estão próximos aos adjetivos. Por exemplo, a sentença "o hotel é caro, desconfortável e mal localizado" gera três tuplas: (hotel, caro), (hotel, desconfortável) e (hotel, mal localizado). Esse padrão permite uma maior cobertura das palavras opinativas referentes a uma única característica. Vale ressaltar que o padrão 2 também reconhece sentenças nas quais as palavras opinativas são ligadas a uma característica por um verbo, como por exemplo, "a piscina estava limpa e divertida".

O padrão 3 por sua vez coleta as características e suas palavras opinativas quando aparecem em ordem inversa no comentário. "Péssimo atendimento", "ótimo quarto" e "perfeita localização" são exemplos de captura do padrão 3.

[2] Termo usado na linguística para descrever o processo de redução (flexionado ou derivado) da palavra.
[3] Disponível em https://goo.gl/SAe9Rr

O padrão 4 verifica se junto das características se encontram verbos e advérbios. A identificação de verbos é importante pois segundo Klavans e Kan (1998), eles também transmitem sentimentos que podem ser explorados. Foi verificado na análise manual que a maioria das tuplas que são encontradas nesse padrão contém inicialmente o advérbio "não". Por exemplo: "não gostei dos quartos" e "não gostei da localização".

O padrão 5 contempla o caso do verbo vir antes do advérbio. A maioria das tuplas identificadas por esse padrão tem o verbo "gostar", como em "gostei muito da comodidade" e "gostei demais do atendimento".

A cobertura dos padrões linguísticos propostos nesta Dissertação é mostrada na Tabela 5. É importante ressaltar que dos 9.909 comentários presentes no Córpus anotado, 8.066 comentários foram atingidos por pelo menos um dos padrões linguísticos utilizados, cobrindo, portanto, 81,41% dos comentários. Além disso, a maior ocorrência dos padrões linguísticos encontrados nos comentários foi 22 vezes em apenas um comentário, enquanto a menor foi 0 vezes em 1.843 comentários.

Tabela 5 – Ocorrência dos padrões linguísticos propostos

Padrão	Quantidade
1	17.882
2	491
3	6.057
4	263
5	242

Verifica-se na Tabela 5 que o padrão 1 foi identificado 17.882 vezes, sendo maior que o somatório da ocorrência dos outros padrões, mostrando a estrutura "sujeito + verbo de ligação + predicativo" geralmente identifica comentários opinativos.

A Tabela 6 mostra quantos comentários foram atingidos por cada padrão. O padrão 1 foi o que mais atingiu comentários do Córpus, com 77,2%, enquanto o padrão 5 foi encontrado em apenas 1,9% dos comentários presentes no Córpus.

Tabela 6 – Quantidade de comentários atingidos por cada padrão

Padrão	Quantidade	Taxa de cobertura
1	7.653	77,2%
2	432	4,3%
3	3.663	36,9%
4	236	2,3%
5	196	1,9%

3.2.2 Adaptações na variável riqueza do vocabulário

Sousa (2015) propôs que a variável riqueza do vocabulário fosse definida pela corretude do comentário, ou seja, quanto mais correto um comentário fosse escrito, mais útil seria a opinião. Nesta Dissertação, além de considerarmos a corretude do comentário, é proposta a utilização de mais três variáveis que medem a riqueza léxica do texto: o índice TTR (do inglês: *type-token ratio*) (TEMPLIN, 1957), o índice de Maas (MASS, 1972) e o índice MTLD (do inglês: *Measure of Textual Lexical Diversity*) (MCCARTHY, 1993).

Torruella e Capsada (2013) afirmam que a riqueza léxica de um texto dá uma ideia do número de diferentes termos usados e a diversidade do vocabulário. Além disso, define três classes de índices que calculam a riqueza.

A primeira classe de índices é baseada na relação direta entre o número de termos e de palavras. Nesta classe se encontra o índice TTR e é definido pela Equação (3.1)

$$TTR = \frac{t}{n} \tag{3.1}$$

onde t é o número de termos únicos e n é o número de palavras.

A segunda classe de índices tem sido desenvolvida baseada em logaritmos. A justificativa para tal é que a função cresce de tal forma a se adaptar melhor ao comportamento da relação existente entre os termos únicos e o número de palavras no texto. Aqui se encontra o índice de Maas, definido pela Equação (3.2)

$$Maas = \frac{\log n - \log t}{(\log n)^2} \tag{3.2}$$

onde, mais uma vez, t é o número de termos únicos e n é o número de palavras.

Ainda existe uma terceira classe de índices, obtidos a partir de cálculos mais complexos, onde se encontra o índice MTLD. O cálculo do índice MTLD é realizado dividindo o texto em segmentos e o índice TTR é calculado para cada segmento. O tamanho de cada segmento s é variável e depende precisamente do valor do índice TTR, sendo o segmento incrementado até que o valor do índice TTR alcance 0,72. Um exemplo de McCarthy e Jarvis (2010), considerando o texto "*...of the people by the people for the people...*" , os valores do índice TTR para cada palavra seriam: *of* (1,00) *the* (1,00) *people* (1,00) *by* (1,00) *the* (0,80) *people* (0,667) *for* (0,714) *the* (0,625) *people* (0,556). Quando o valor de TTR alcança 0,72, o contador de segmentos incrementa em 1 e o valor do índice TTR é zerado. Desta forma, seguindo o exemplo anterior, o cálculo do índice MTLD é feito da seguinte maneira: *of* (1,00) *the* (1,00) *people* (1,00) *by* (1,00) *the* (0,80) *people* (0,667) $\parallel s = s + 1 \parallel$ *for* (1,00) *the* (1,00) *people* (1,00) ... e assim por diante. Considera-se que um segmento é completo quando o índice TTR da última palavra é 0,72 (MCCARTHY; JARVIS, 2010). Quando o índice TTR da última palavra é diferente de 0,72 considera-se

calcular o segmento parcial do texto (SP), por meio da Equação (3.3) (LU, 2014)

$$SP = \frac{1 - TTR_{lastword}}{1 - 0,72} \tag{3.3}$$

onde $TTR_{lastword}$ é o valor do índice TTR na última palavra do texto. Por exemplo, se o texto contém 4 segmentos completos e o índice TTR da última palavra o texto é 0,887, o segmento parcial seria 0,404 e ao final, a quantidade de segmentos seria 4 + 0,404 = 4,404. A justificativa para este cálculo adicional é que uma pequena seção do texto, em termos de *tokens*, sempre terá o índice TTR alto, o que causaria inconstância na diversidade léxica do texto por completo. Ao final do texto, são contados em quantos segmentos foram divididos o texto e o valor do índice MTLD é calculado pela Equação (3.4)

$$MTLD = \frac{n}{s + SP} \tag{3.4}$$

onde n é o tamanho do texto em número de palavras, s é o número de segmentos completos e SP é a quantidade de segmentos parciais. Uma nova execução do índice MTLD é feita com o texto sendo processado de forma inversa. A média dos dois valores é o valor final do índice MTLD (MCCARTHY; JARVIS, 2010).

Em relação ao tamanho do texto, Torruella e Capsada (2013) comprovaram que os índices pertencentes à primeira classe são sensíveis ao tamanho do texto enquanto os índices pertencentes à segunda e terceira classes não são. Além disso, o índice de Maas mostrou ser o mais estável em respeito ao tamanho do texto.

A utilização dos índices é justificada pela sua capacidade de detectar e quantificar as diferenças na riqueza léxica entre diferentes comentários. Quanto mais correto e rico de vocabulário for o comentário, mais importante ele se torna. Dessa forma, para mapear a riqueza do vocabulário deve-se considerar as quatro variáveis descritas: i) a corretude; ii) a riqueza léxica medida pelo índice TTR; iii) a riqueza léxica medida pelo índice de Maas; e iv) a riqueza léxica medida pelo índice MTLD.

Para o cálculo da variável de corretude das palavras do comentário, foi utilizado um dicionário de palavras retirado da *WordNet.BR* contendo 250.196 palavras. O cálculo das variáveis referentes aos índices TTR, Maas e MTLD foi feito obtendo os valores da quantidade de termos diferentes que são expressos no comentário e o número de palavras existentes no comentário por meio do método de tokenização e contagem de palavras na biblioteca NLTK.

Pelo fato da adaptação da variável riqueza do vocabulário ser baseada na abordagem de Sousa (2015), as quatro variáveis propostas nesta Dissertação para definir a riqueza do vocabulário devem ser convertidas em apenas um valor de entrada. A justificativa para tal vem de Sousa (2015) que em seu trabalho comenta que não poderiam ser criadas muitas variáveis linguísticas, pois ajudaria a aumentar a quantidade de regras do Sistema *Fuzzy*.

A solução proposta foi a implementação de uma RNA MLP que recebe como entrada as quatro variáveis descrita nesta seção e retorna como saída um valor real entre 0 e 3 que definem o quão bom é a riqueza do vocabulário do comentário analisado. A RNA é ilustrada na Figura 9.

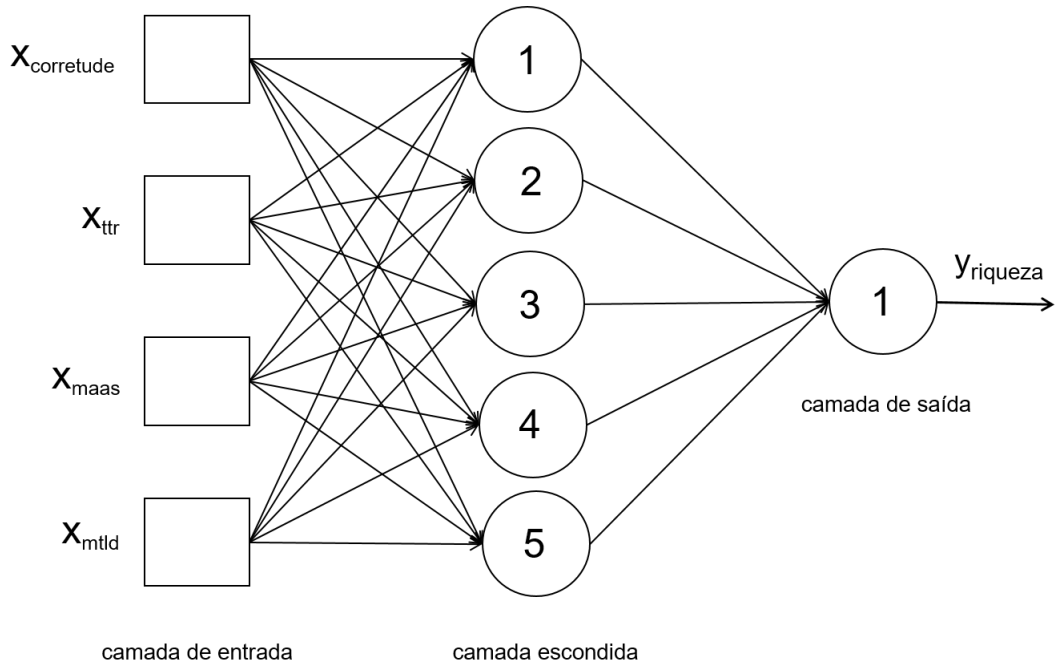


Figura 9 – Topologia: RNA riqueza do vocabulário

A topologia da RNA foi definida por meio do *software* SPSS (IBM Corp., 2011). A RNA contém quatro entradas, cinco neurônios na camada escondida e um neurônio na camada de saída. A função de ativação utilizada na camada escondida foi a tangente hiperbólica e na camada de saída foi a *softmax*. A função do erro utilizada foi a entropia cruzada (*cross-entropy*) (MAGOULAS; VRAHATIS; ANDROULAKIS, 1999).

Para a realização do treinamento e teste foram considerados 370 comentários[4] que foram manualmente classificados como Ruim (RM), Médio (MD), Bom (BM) e Excelente (EXC) com referência aos valores de corretude, TTR, Maas e MTLD. Estes comentários foram utilizados no treinamento da rede, considerando os valores 0, 1, 2 e 3 para as classes Ruim, Médio, Bom e Excelente, respectivamente. A matriz de confusão relacionada ao treinamento da rede é mostrada na Tabela 7. A porcentagem de acertos das probabilidades nas classes realizada na fase de testes da Rede Neural Artificial proposta para a variável riqueza do vocabulário foi de 82,5%.

Por fim, destaca-se que o resultado da RNA serve como entrada para a abordagem utilizando Sistema *Fuzzy* a ser explicada na Seção 3.3 e como base para a abordagem utilizando RNA a ser discutida na Seção 3.4.

[4] Número de comentários definido a partir de uma análise estatística, mostradas nas Equações (1.4) e (1.5) na Seção 1.3.3.

Tabela 7 – Matriz de confusão: RNA riqueza do vocabulário

Valor Real	Valor Predito				Total
	EXC	**BM**	**MD**	**RM**	
EXC	53	12	1	0	**66**
BM	11	85	25	0	**121**
MD	0	23	140	1	**164**
RM	0	0	15	4	**19**
Total	**64**	**120**	**181**	**5**	370

3.3 Abordagem com Sistema *Fuzzy*

Levando em conta as três variáveis de entrada propostas por Sousa (2015) e as adaptações propostas em duas delas, quantidade de tuplas e riqueza do vocabulário, discutidas nas seções 3.2.1 e 3.2.2, o esquema geral da abordagem com Sistema *Fuzzy* é apresentado na Figura 10.

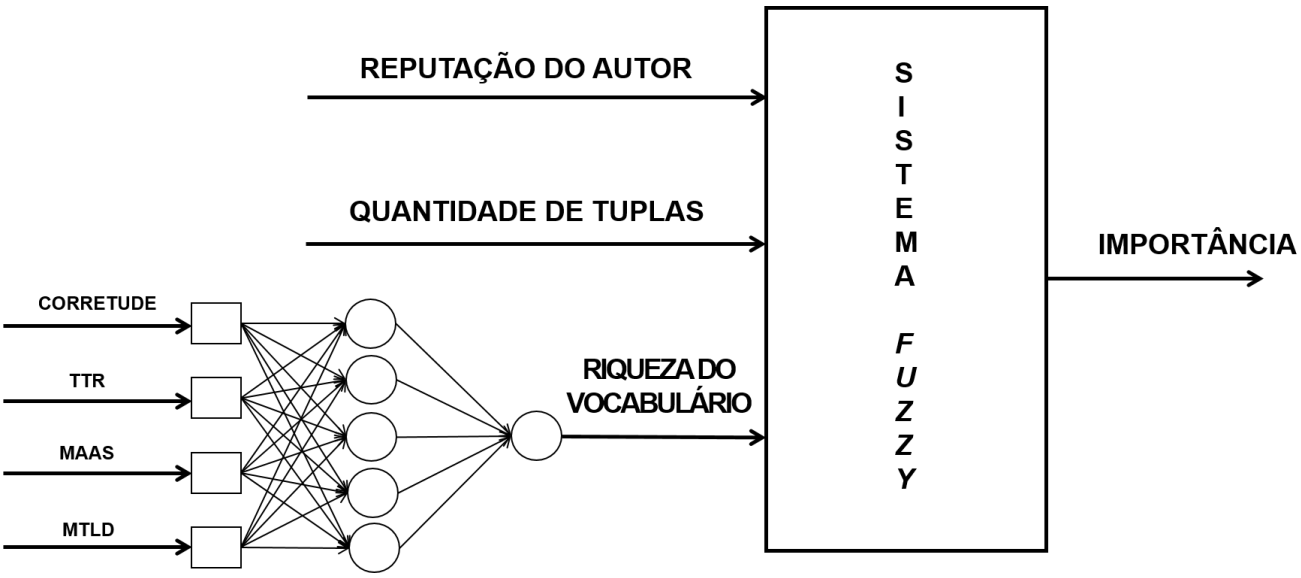


Figura 10 – Abordagem com Sistema *Fuzzy*

A abordagem utiliza o modelo de inferência de Mandami (MAMDANI; ASSILIAN, 1975). A utilização desse modelo é justificada pelo fato do uso de variáveis linguísticas tanto nas entradas quanto na saída do sistema de inferência, o que torna o processo de modelagem do sistema mais intuitivo (SOUSA, 2015).

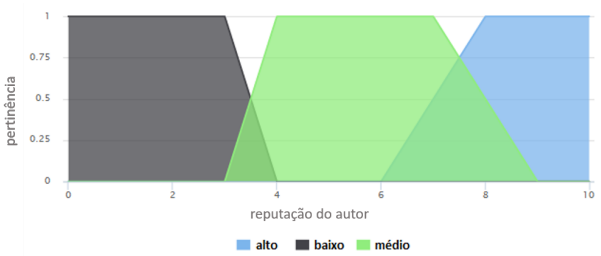
Como explicado na Seção 1.2.1, o sistema de inferência *Fuzzy* contém três etapas: fuzzificação, processo de inferência *Fuzzy* que se baseia na definição das regras e em suas inferências e defuzzificação. A primeira tarefa realizada foi a definição dos valores linguísticos para cada uma das variáveis do sistema, tanto as variáveis de entrada quanto a variável de saída. Os valores linguísticos de cada variável são mostrados na Tabela 8

Tabela 8 – Valores linguísticos das variáveis do sistema de inferência *Fuzzy*

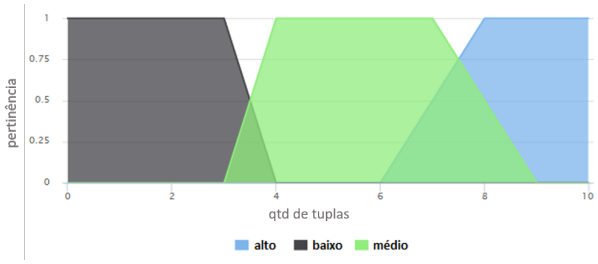
Variável	Valores Linguísticos
Reputação do Autor	Baixo (BX), Médio (MD), Alto (AL)
Quantidade de Tuplas	Baixo (BX), Médio (MD), Alto (AL)
Riqueza do Vocabulário	Ruim (RM), Médio (MD), Bom (BM), Excelente (EXC)
Importância	Insuficiente (ISF), Suficiente (SF), Bom (BM), Excelente (EXC)

Os valores linguísticos de todas as variáveis foram baseados em Sousa (2015). Foi necessário adaptar a variável riqueza do vocabulário pois, como explicado na Seção 3.2.2, o valor resultante provém de uma RNA que inclui quatro entradas e retorna uma saída que está entre 0 e 3, mapeando cada uma das saídas numéricas a uma variável linguística.

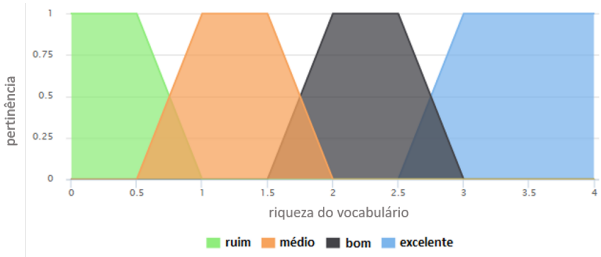
Na etapa de fuzzificação obtém-se o grau de pertinência com que cada entrada pertence a cada conjunto *Fuzzy*. Cada uma das entradas foi previamente limitada no universo de discurso e associada a um grau de pertinência em cada conjunto *Fuzzy* por meio do conhecimento do especialista. A obtenção do grau de pertinência se dá pela análise das funções de pertinência envolvidas no sistema. As funções de pertinência da abordagem proposta são mostradas na Figura 11: reputação do autor (Fig 11a), quantidade de tuplas (Fig 11b), riqueza do vocabulário (Fig 11c) e da importância do comentário (Fig 11d). É importante mencionar que estas funções de pertinência não são similares às funções de pertinência propostas por Sousa (2015), pois houveram alterações no comportamento da variável riqueza do vocabulário, fazendo com que as funções de pertinência fossem revisadas.



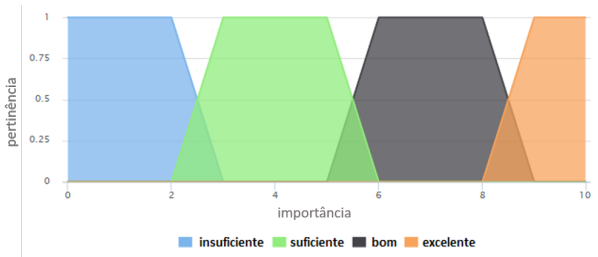
(a) Funções de pertinência: reputação do autor



(b) Funções de pertinência: quantidade de tuplas



(c) Funções de pertinência: riqueza do vocabulário



(d) Funções de pertinência: importância

Figura 11 – Funções de pertinência

A etapa referente ao processo de inferência recebe as entradas fuzzificadas e aplica-

as de acordo com cada regra contida na base de regras. A base de regras contém as regras *Fuzzy* que representam o conhecimento do especialista ao processo/sistema. A estrutura de uma regra é: **SE** $(v_{e1} = a)$ **E** $(v_{e2} = b)$ **E** $(v_{e3} = c)$ **ENTÃO** $(v_s = d)$, onde v_{e1}, v_{e2} e v_{e3} são as variáveis de entrada e v_s é a variável de saída. A base de regras proposta para a abordagem com Sistema *Fuzzy* é mostrada na Tabela 9. Mais uma vez é válido ressaltar que a base de regras é diferente da abordagem TOP(X) original devido as adaptações na variável riqueza de vocabulário, sendo preciso assim uma revisão.

Tabela 9 – Base de Regras do Sistema *Fuzzy*

	Quantidade de Tuplas/Riqueza do Vocabulário											
Autor	BX/RM	BX/MD	BX/BM	BX/EXC	MD/RM	MD/MD	MD/BM	MD/EXC	AL/RM	AL/MD	AL/BM	AL/EXC
BX	ISF	ISF	ISF	SF	ISF	SF	SF	BM	SF	SF	BM	BM
MD	ISF	ISF	ISF	SF	SF	SF	BM	BM	SF	SF	BM	EXC
AL	SF	SF	SF	BM	SF	BM	BM	EXC	BM	BM	EXC	EXC

Como exemplo de leitura das regras da Tabela 9, utiliza-se os valores referentes à intersecção da primeira linha, referente ao autor e da primeira coluna, referente à quantidade de tuplas e riqueza do vocabulário. As entradas são: Autor = *baixo* (BX), Quantidade de tuplas = *baixo* (BX) e Riqueza do vocabulário = *ruim* (RM), resultando na saída Importância = *insuficiente* (ISF).

A etapa final do sistema de inferência *Fuzzy* é a defuzzificação. Para obter um valor numérico para a variável de saída importância do comentário foi utilizado o método centro de área (centro de gravidade[5] ou centróide). Este método encontra a média aritmética entre os centros de gravidade dos conjuntos *Fuzzy* aos quais o elemento pertence, ponderados pelo grau de pertinência. Este é o método mais complexo e portanto mais demorado, mas por outro lado é o método mais preciso (WEBER; KLEIN, 2003). Sua equação matemática é definida pela Equação (3.5)

$$Y = \frac{\sum_{i=1}^{qc} \mu_i \cdot Y_i}{\sum_{i=1}^{qc} \mu_i} \tag{3.5}$$

onde Y é o valor numérico resultante, qc é a quantidade de conjuntos *Fuzzy* a que o elemento pertence, μ_i é o grau de pertinência com que o elemento pertence ao i-ésimo conjunto *Fuzzy* e Y_i é o centro de gravidade do i-ésimo conjunto *Fuzzy* a que o elemento pertence (SANTOS et al., 2014). A execução da abordagem foi realizada na ferramenta Athena e o arquivo controlador de lógica *Fuzzy* (arquivo `.fcl`) se encontra no Apêndice B.

[5] Na maioria dos casos, o centro da área está na mesma posição do centro de gravidade, então estes nomes frequentemente denotam o mesmo método (WEBER; KLEIN, 2003).

3.4 Abordagem com Rede Neural Artificial

No decorrer da seção 1.2.2 foram abordados os principais pontos de uma Rede Neural Artificial. A escolha de uma RNA se deu devido ao seu poder na resolução de problemas que envolvem classificação de padrões. A definição da abordagem se deu por três etapas: i) definição da arquitetura da RNA; ii) definição da topologia da RNA; e iii) treinamento e teste da RNA.

Na etapa de definição da arquitetura da RNA foram estudados dois tipos de redes que resolvem problemas envolvendo classificação de padrões: Perceptron Multi-camadas (MLP) utilizando o algoritmo de treinamento *backpropagation* e Redes de Base Radial (RBF).

No treinamento das redes MLP com *backpropagation*, duas fases são observadas: a "propagação adiante" (*forward*) na qual os sinais de entrada de uma amostra são propagados camada a camada até a produção das respectivas saídas, levando-se em consideração apenas os valores atuais de pesos sinápticos e limiares, calculando o erro a partir das respostas desejadas e a "propagação reversa" (*backward*), que leva em consideração esse erro e ocorre a atualização dos pesos sinápticos iniciando da saída até a camada de entrada. As aplicações sucessivas das fases *forward* e *backward* fazem com que os pesos sinápticos e limiares dos neurônios se ajustem automaticamente em cada iteração, implicando na gradativa diminuição do erro produzido pelas respostas da rede.

Já o treinamento das redes RBF, diferentemente das redes MLP, é constituída de dois estágios bem distintos entre si. O primeiro estágio é associado com o ajuste dos pesos na única[6] camada escondida existente na rede, adotando um método de aprendizagem auto-organizado, ou seja, não-supervisionado, dependente apenas das características dos dados de entrada. Neste primeiro estágio, a função de ativação é composta por funções de base radial como as gaussianas. Apenas quando é finalizado o primeiro estágio de treinamento inicia-se o segundo estágio, que utiliza o mesmo método de aprendizagem supervisionado na camada de saída da rede utilizado nas redes MLP. É importante destacar que não há o *backpropagation* da rede inteira, apenas dos neurônios da camada de saída das redes RBF (MAK; ALLEN; SEXTON, 1993; KRIESEL, 2007; SILVA; SPATTI; FLAUZINO, 2010).

Apesar da aplicabilidade de ambas as arquiteturas para resolverem o problema em questão desta Dissertação – classificação dos comentários em graus de importância, uma rede RBF pode requerer mais neurônios em sua camada intermediária quando comparada com as redes MLP (SILVA; SPATTI; FLAUZINO, 2010). Por outro lado, o treinamento das redes RBF geralmente é mais rápido que o treinamento das redes MLP (CHEN; COWAN; GRANT, 1991).

[6] É comum associar a rede RBF a uma rede com uma camada escondida, embora redes RBF com mais de uma camada escondida tenham sido propostas (HE; LAPEDES, 1994; BRAGA; CARVALHO; LUDERMIR, 2007)

Foram considerados 370 comentários, classificados manualmente em Excelentes, Bons, Suficientes e Insuficientes. Esta foi a analogia referente ao Sistema *Fuzzy*, uma vez que as variáveis linguísticas da abordagem com Sistema *Fuzzy* proposta nesta Dissertação resultam nesta classificação e por este motivo a saída da RNA deve seguir o mesmo padrão.

A próxima etapa consistiu em definir as topologias a serem usadas. Ressalta-se que toda a tarefa de definição da arquitetura, topologia (quantidade de neurônios na camada escondida, funções de ativação e funções de erro) e execução da RNA foi realizada no *software* SPSS e de acordo com a documentação do *software*, a definição da topologia referente às redes MLP é feita baseada em um algoritmo que determina o "melhor" número de neurônios da camada escondida, definido por *Expert Architecture Selection*[7].

Basicamente, o algoritmo retira uma amostra aleatória de todo o conjunto de dados (comentários) e os divide em dois subconjuntos: treinamento (70% dos dados) e teste (30% dos dados) de tamanho $NC = min(1000, memsize)$, onde $memsize$ é o tamanho do conjunto de dados armazenados na memória. O número de neurônios da camada escondida é definido por meio de testes de várias redes treinadas com o conjunto de dados que atingem o erro de teste mínimo definido pelo algoritmo. Este algoritmo é baseado na técnica de validação cruzada (KOHAVI, 1995), cujo propósito é avaliar a aptidão de cada topologia candidata quando são aplicadas a um conjunto de dados diferente do utilizado no ajuste dos seus parâmetros internos.

A topologia da RNA com o tipo MLP é mostrada na Figura 12 e consiste de seis variáveis na camada de entrada, seis neurônios na camada escondida e um neurônio na camada de saída, representando cada classe referente à importância do comentário. A função de ativação utilizada na camada escondida foi a tangente hiperbólica e na camada de saída usou-se a função *softmax*. A função do erro usada nesta topologia foi a entropia cruzada (*cross-entropy*).

A definição da topologia referente às redes RBF seguiu praticamente os mesmos procedimentos utilizados na definição da topologia das redes MLP. O *software* SPSS usa um algoritmo que determina o "melhor" número de neurônios na camada escondida da rede RBF, definido por *Automatic Selection of Number of Basis Functions*[8] e calcula automaticamente os valores mínimo e máximo de um intervalo definido pelo usuário e encontra o melhor número de neurônios dentro desse intervalo.

A topologia da RNA com o tipo RBF é ilustrada na Figura 13, e consiste de seis variáveis na camada de entrada, oito neurônios na camada escondida e um neurônio na camada de saída. A função de ativação usada na camada escondida foi a *softmax* e a função

[7] Disponível em https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_mlp_architecture_expert.htm

[8] Disponível em https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_rbf_training_autoselect_numhiddenunits.htm

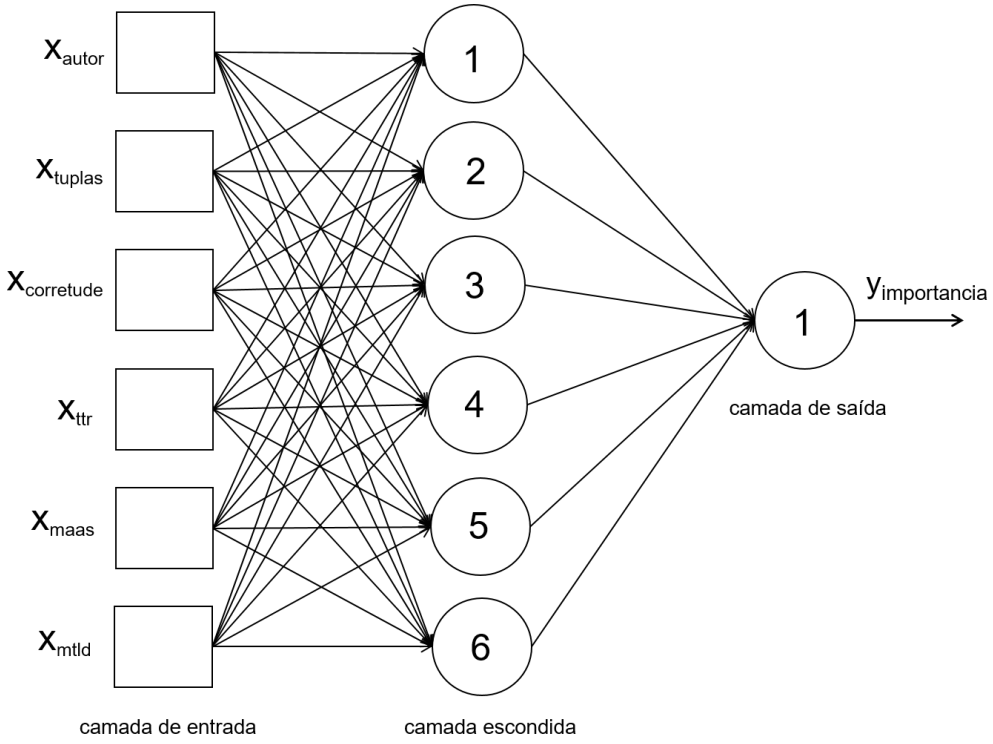


Figura 12 – Topologia: RNA *Multi-Layer Perceptron*

aplicada na camada de saída foi a identidade. A justificativa por utilizar a função *softmax* na camada escondida se dá pelo fato do treinamento no primeiro estágio usar apenas funções de base radial, chamada neste caso de função de base radial normalizada (do inglês: *normalized radial basis function*, NRBF) (BUGMANN, 1998; HEIMES; HEUVELN, 1998), a qual utiliza a função *softmax* (DUCH; JANKOWSKI, 1999). A função do erro utilizada nesta topologia foi a soma dos erros quadrados (DRAPER; SMITH, 2014).

Vale ressaltar que as variáveis que serviram de entrada para ambas as topologias das RNAs tipo MLP e RBF foram: i) reputação do autor; ii) quantidade de tuplas; iii) corretude do comentário; iv) índice TTR; v) índice Maas; e vi) índice MTLD. Diferente do que proposto no Sistema *Fuzzy* explicado na Seção 3.3, a abordagem utilizando RNA suporta muito bem todas as variáveis sendo consideradas juntas, facilitando a análise final e evitando que uma das entradas da RNA seja uma outra RNA, assim como é uma das entradas do Sistema *Fuzzy* proposto.

Finalmente, foi possível verificar que a RNA com o tipo RBF contém mais neurônios na camada escondida que a RNA com o tipo MLP (oito e seis, respectivamente), corroborando com o que foi explicado durante este capítulo. No entanto, o tempo de treinamento da topologia contendo as redes RBF (0,495 segundos) foi um pouco mais lento que as redes MLP (0,105 segundos). A porcentagem de acertos das probabilidades na fase de teste das redes RBF e redes MLP foram 70,3% e 80%, respectivamente.

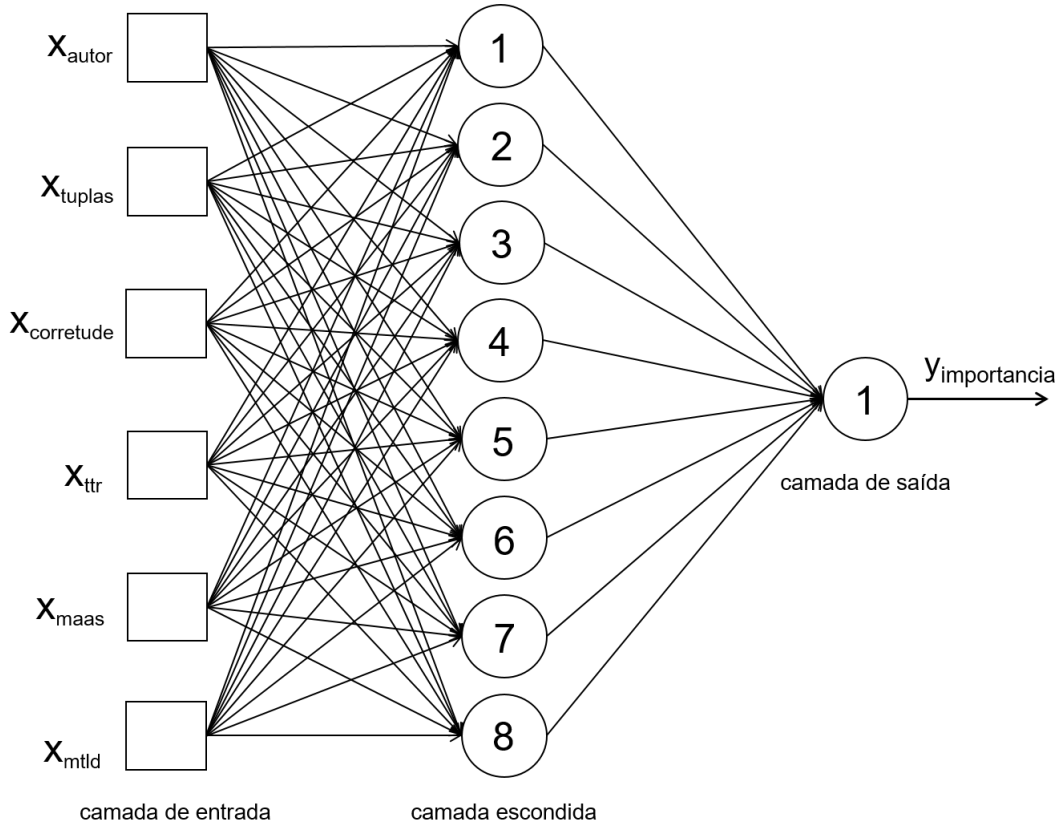


Figura 13 – Topologia: Redes de Base Radial

3.5 Considerações Finais

Este Capítulo apresentou detalhes das abordagens propostas nesta Dissertação que utilizam Sistema *Fuzzy* e Redes Neurais Artificiais, com o objetivo de realizar um estudo comparativo entre elas. Também foram detalhadas as adaptações nas variáveis de entrada quantidade de tuplas e riqueza do vocabulário implementadas nesta versão.

As adaptações referentes à quantidade de tuplas utilizam uma base de características para limitar as características a serem analisadas, além de novos padrões linguísticos propostos baseados nos existentes na literatura para o Português. Em relação às adaptações realizadas na variável riqueza do vocabulário, três índices que medem a riqueza léxica de um texto foram adicionados à corretude do texto para medir tal riqueza. Para este fim, foi necessária uma RNA que considera as quatro medidas e retorna um valor que significa a riqueza léxica daquele comentário levando em conta os quatro índices.

Tais adaptações foram utilizadas nas duas abordagens implementadas. A abordagem utilizando Sistema *Fuzzy* contém três entradas: i) reputação do autor; ii) quantidade de tuplas; e iii) riqueza do vocabulário, sendo esta última a saída de uma RNA contendo quatro índices de riqueza léxica. A abordagem usando RNA teve dois tipos de redes que resolvem o problema de classificação de padrões: *Multi-Layer Perceptron* (MLP) e Função de Base Radial (RBF). Ambas as redes contém seis entradas: i) reputação do autor; ii)

quantidade de tuplas; iii) corretude do comentário; iv) índice TTR; v) índice Maas; e vi) índice MTLD, uma camada escondida com seis e oito neurônios, respectivamente e uma camada de saída com um neurônio, representando cada classe que diz sobre a importância: insuficiente, suficiente, bom e excelente. Os experimentos realizados e os resultados são discutidos no próximo Capítulo.

4 Experimentos e Discussões

Neste Capítulo são discutidos os resultados referentes às abordagens propostas nesta Dissertação. Adicionalmente, são analisados os modelos computacionais baseados em Redes Neurais Artificiais tipo MLP e RBF, levando em consideração a importância das variáveis de entrada em relação à saída de cada uma das redes. Ao final, a comparação dos modelos computacionais é realizada, junto com a abordagem base.

4.1 Resultados

Com o objetivo de comparar as abordagens, foi realizado um experimento com o subcórpus de importância. Para cada comentário, utilizou-se os mesmos métodos para extrair as informações das variáveis de entrada de cada abordagem. Os resultados são apresentados por meio de uma matriz de confusão, contendo o número de classificações corretas em oposição às classificações preditas para cada classe, a saber: Excelente (EXC), Bom (BM), Suficiente (SF) e Insuficiente (ISF). Para avaliação dos modelos, calculou-se as medidas de precisão (P), cobertura (R) e medida-F (F) para cada classe, bem como a taxa de erro. Calculou-se também a precisão total da abordagem, que representa a acurácia do modelo. É importante destacar que essas medidas são normalmente usadas em avaliação de abordagens na área de aprendizagem de máquina (POWERS, 2011).

A Tabela 10 mostra a matriz de confusão para a abordagem TOP(X) original, utilizada como base no processo de comparação. As medidas de precisão, cobertura, medida-F e taxa de erro por classe são apresentadas na Tabela 11.

Tabela 10 – Matriz de confusão: abordagem TOP(X) original

Valor Real	Valor Predito				Total
	EXC	**BM**	**SF**	**ISF**	
EXC	16	2	2	0	**20**
BM	42	18	33	5	**98**
SF	9	26	58	30	**123**
ISF	2	1	23	103	**129**
Total	**69**	**47**	**116**	**138**	370

Destaca-se que a precisão é calculada como a porcentagem de exemplos corretamente classificados como positivos em cada classe, por meio da fórmula

$$P = \frac{TP}{(TP + FP)} \tag{4.1}$$

Tabela 11 – Medidas de avaliação: abordagem TOP(X) original

Classe	**P**	**R**	**F**	**Taxa de Erro**
EXC	23,18%	80%	35,95%	20%
BM	38,29%	18,36%	24,82%	81,64%
SF	50%	47,15%	48,53%	52,85%
ISF	74,63%	79,84%	77,15%	20,16%

onde TP representa os comentários classificados corretamente (do inglês: *true positive*) e FP representa os comentários classificados incorretamente (do inglês: *false positive*). Por exemplo, para a classe EXC, 16 comentários foram classificados corretamente, mas 53 comentários foram classificados incorretamente, sendo a precisão igual a 16/(16+53) = 23,18%. Para a classe BM, 18 comentários foram classificados corretamente (TP), mas 29 comentários foram classificados incorretamente (FP), alcançando a precisão igual a 18/(18+29) = 38,29%.

A cobertura é obtida por meio da porcentagem de exemplos corretamente classificados como positivos em relação ao total de instâncias da classe do subcórpus de importância, por meio da fórmula

$$R = \frac{TP}{(TP + FN)} \tag{4.2}$$

onde TP representa os comentários classificados corretamente e FN representa os comentários classificados incorretamente em relação à análise do subcórpus de importância (do inglês: *false negative*). Por exemplo, para a classe EXC, a cobertura é igual a 16/(16+4) = 80%, pois 16 comentários foram classificados corretamente (TP), mas 4 comentários foram classificados incorretamente (FN).

A medida-F é uma média harmônica entre precisão e cobertura derivada de Rijsbergen (1979). O cálculo da medida-F é realizado segundo a equação

$$F = \frac{2 * P * R}{P + R} \tag{4.3}$$

onde P representa o valor da precisão e R representa o valor da cobertura. Existem formas de ponderar os valores de precisão e cobertura de acordo com o objetivo do cálculo (PISKORSKI; YANGARBER, 2013), mas nesta Dissertação não foi necessário ponderar, pois o intuito da utilização da medida-F foi avaliar por completa a abordagem em sua precisão e cobertura em cada classe. No caso da classe EXC, a medida-F é igual a 35,95% e para a classe BM a medida-F é igual a 24,82%.

A taxa de erro é calculada como o número de comentários classificados de forma incorreta dividido pelo total de comentários pertencentes à classe. Por exemplo, para a classe EXC, a taxa de erro é igual a 4/20 = 20%.

Por fim, a precisão total da abordagem (acurácia) é calculada como o número de exemplos classificados corretamente em cada classe pelo total geral da amostra. No caso, tem-se 16+18+58+103 = 195 comentários corretamente classificados. Logo, a precisão total da abordagem base é igual a 195/370 = 52,7%.

A Tabela 12 apresenta a matriz de confusão para a abordagem com Sistema *Fuzzy* e adaptações e a Tabela 13 mostra as medidas de precisão, cobertura, medida-F e taxa de erro por classe. A precisão total da abordagem foi de 60,54%.

Tabela 12 – Matriz de confusão: abordagem com Sistema *Fuzzy* e adaptações

Valor Real	Valor Predito				Total
	EXC	**BM**	**SF**	**ISF**	
EXC	13	5	2	0	**20**
BM	1	44	38	15	**98**
SF	0	11	56	56	**123**
ISF	0	1	17	111	**129**
Total	**14**	**61**	**113**	**182**	370

Tabela 13 – Medidas de avaliação: abordagem com Sistema *Fuzzy* e adaptações

Classe	**P**	**R**	**F**	**Taxa de Erro**
EXC	92,8%	65%	76,4%	35%
BM	72,1%	44,8%	55,3%	55,2%
SF	49,5%	45,5%	47,4%	54,5%
ISF	60,9%	86%	71,3%	14%

A matriz de confusão referente à abordagem com RNA tipo MLP é mostrada na Tabela 14 e na Tabela 15 são apresentadas as medidas de precisão, cobertura, medida-F e taxa de erro por cada classe. A precisão total desta abordagem em questão foi de 71,08%.

Tabela 14 – Matriz de confusão: abordagem com RNA tipo MLP

Valor Real	Valor Predito				Total
	EXC	**BM**	**SF**	**ISF**	
EXC	15	5	0	0	**20**
BM	7	67	23	1	**98**
SF	0	21	80	22	**123**
ISF	0	1	27	101	**129**
Total	**22**	**94**	**130**	**124**	370

Tabela 15 – Medidas de avaliação: abordagem com RNA tipo MLP

Classe	**P**	**R**	**F**	**Taxa de Erro**
EXC	68,1%	75%	71,4%	25%
BM	71,2%	68,3%	69,7%	31,7%
SF	61,5%	65%	63,2%	35%
ISF	81,4%	78,2%	79,8%	21,8%

Na Tabela 16 é apresentada a matriz de confusão referente à abordagem com RNA tipo RBF, enquanto que as medidas de precisão, cobertura, medida-F e taxa de erro por classe são mostradas na Tabela 17. Sua precisão total foi de 64,32%.

Tabela 16 – Matriz de confusão: abordagem com RNA tipo RBF

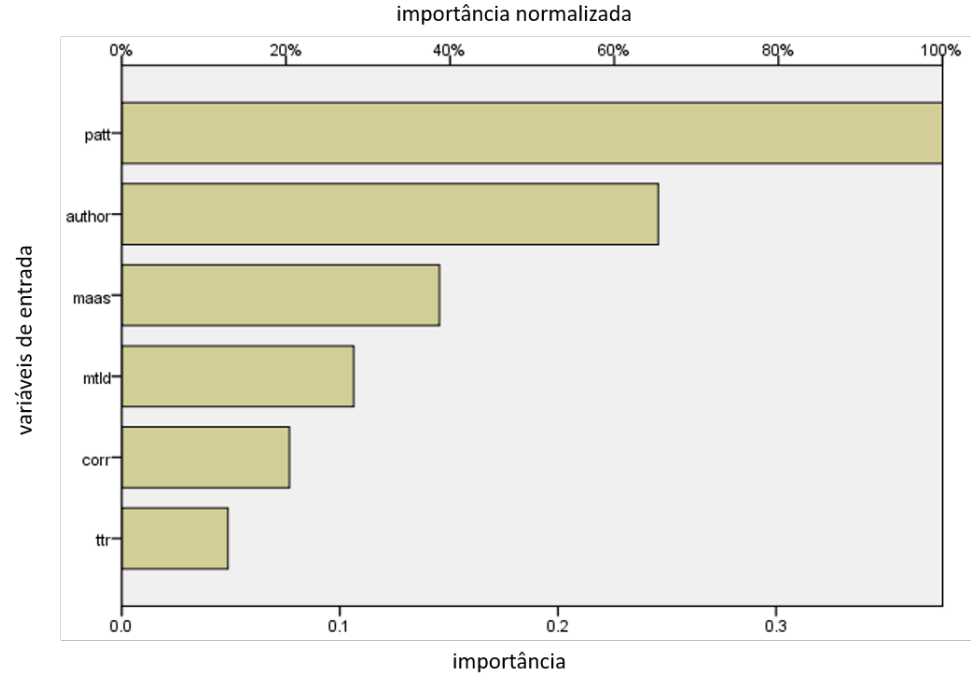
Valor Real	Valor Predito				Total
	EXC	**BM**	**SF**	**ISF**	
EXC	14	5	0	1	**20**
BM	3	64	26	5	**98**
SF	1	26	59	37	**123**
ISF	0	2	26	101	**129**
Total	**18**	**97**	**111**	**144**	370

Tabela 17 – Medidas de avaliação: abordagem com RNA tipo RBF

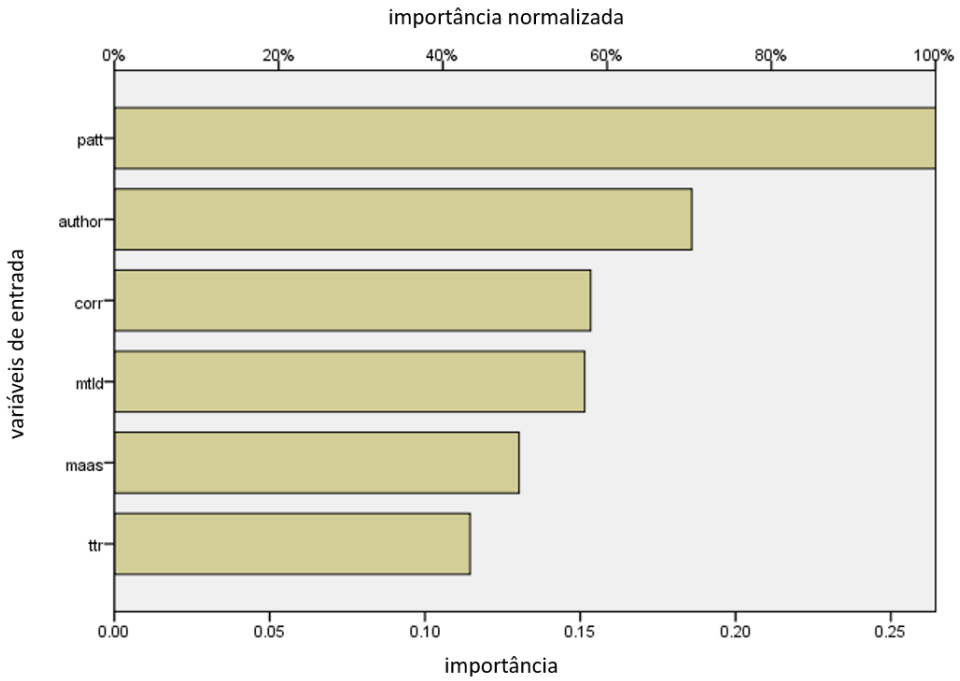
Classe	**P**	**R**	**F**	**Taxa de Erro**
EXC	77,7%	70%	73%	30%
BM	65,9%	65,3%	65,6%	34,7%
SF	53,1%	47,9%	50,4%	52,1%
ISF	70,1%	78,2%	73,9%	21,8%

4.2 Análise dos Modelos RNA

Com relação às duas abordagens que utilizam RNA, analisou-se a importância das variáveis de entrada em relação à saída para descobrir quais características de um comentário afetam a percepção da utilidade dele, ou seja, quais das entradas da RNA têm mais influência no resultado final. O *software* SPSS, ao final da execução do treinamento e teste das RNAs, apresenta um relatório com a Análise de Importância da Variável Independente (do inglês: *Independent Variable Importance Analysis*). Os gráficos para os modelos MLP e RBF são mostrados nas figuras 14a e 14b, respectivamente.



(a) Importância das variáveis: RNA MLP



(b) Importância das variáveis: RNA RBF

Figura 14 – Importância das variáveis referente às abordagens com RNA

Esses resultados são obtidos a partir de uma análise de sensibilidade entre as variáveis de entrada de cada abordagem que investiga a forma como a variação na saída de um modelo numérico pode ser atribuído às variações dos seus fatores de entrada (PIANOSI et al., 2016). Observa-se nas figuras 14a e 14b que as variáveis mais importantes em ambas as abordagens são as referentes à quantidade de tuplas (*patt*) e reputação do autor (*author*). Entende-se que as variáveis referentes à riqueza de vocabulário são as que menos interferem no resultado final da importância segundo o cálculo de sensibilidade, mais especificamente o índice TTR (*ttr*), o qual tem mais dependência do tamanho do texto (ver Seção 3.2.2), portanto, justificada como a entrada mais inconstante das quatro entradas de riqueza de vocabulário.

Para dar mais robustez ao fato da importância das variáveis nas abordagens usando redes neurais artificiais propostas, foram utilizadas técnicas de seleção de características, para serem comparadas com a análise de sensibilidade realizada pelo *software* SPSS. Três técnicas comumente usadas na literatura e de fácil interpretação dos resultados foram utilizadas neste trabalho: i) Seleção Univariante (do inglês: *Univariate Selection*) utilizando o teste qui-quadrado (PEARSON, 1900; YATES, 1934); ii) Eliminação Recursiva de Características (do inglês: *Recursive Feature Elimination*, RFE) (GUYON et al., 2002); e iii) Importância de características utilizando Árvores de Decisão Aleatórias (ADA) (do inglês: *Randomized Decision Trees*).

Cada algoritmo recebeu como entrada os valores das variáveis de entrada e de saída de cada comentário e retornou como saída um *ranking* com as variáveis que são mais importantes de acordo com os parâmetros de cada técnica de seleção de características. Esses *rankings* são mostrados na Tabela 18.

Tabela 18 – Importância das variáveis: por técnicas de seleção de características

Seleção Univariante		**RFE**		**ADA**	
1. Qtd. Tuplas	0,403	**1. Qtd. Tuplas**	*TRUE*	**1. Qtd. Tuplas**	0,304
2. Rep. Autor	0,372	**2. Rep. Autor**	*TRUE*	**2. Rep. Autor**	0,164
3. MTLD	0,208	3. Maas	*FALSE*	3. Maas	0,143
4. Corretude	0,010	4. Corretude	*FALSE*	4. MTLD	0,135
5. TTR	0,004	5. MTLD	*FALSE*	5. Corretude	0,134
6. Maas	0,003	6. TTR	*FALSE*	6. TTR	0,120

Os valores referentes às técnicas de Seleção Univariante e Importância de características utilizando ADA são normalizados, ou seja, a soma dos valores resulta em 1. A técnica RFE seleciona as c características mais importantes e atribui o valor booleano $TRUE$ e atribui o valor booleano $FALSE$ para o restante. Como não havia um *ranking* numérico para definir qual é mais importante, o algoritmo foi executado seis vezes, selecionando sempre o total de $c - 1$ características. Por exemplo, a primeira execução da técnica RFE

selecionou as cinco mais importantes, na qual o índice TTR foi a única variável que teve valor $FALSE$. A segunda execução contou com cinco características e a técnica RFE selecionou as quatro mais importantes, atribuindo o valor $FALSE$ para a variável referente ao índice MTLD. As outras execuções seguiram o mesmo esquema até que as duas últimas características fossem comparadas.

Percebe-se que em todas as técnicas de seleção de características, as variáveis referentes à quantidade de tuplas e reputação de autor foram consideradas importantes para definir a importância dos comentários. Por outro lado, há uma variação de qual variável de riqueza de vocabulário é a menos importante. O índice TTR é o menos importante em três das quatro análises discutidas (incluindo a análise feita pelo SPSS por meio da análise de sensibilidade).

4.3 Comparação dos Resultados

Após os resultados das execuções das abordagens serem apresentados por meio de suas matrizes de confusão, foi possível realizar a comparação entre as abordagens propostas nesta Dissertação. Como as abordagens que utilizam Sistema *Fuzzy* e Redes Neurais Artificiais são paramétricas (RUSSELL; NORVIG, 2009), fez-se uma comparação cuidadosamente planejada. As Seções 3.3 e 3.4 especificam os parâmetros de cada abordagem.

Para facilitar a comparação, a Tabela 19 mostra as precisões totais (acurácia) das quatro abordagens estudadas nesta Dissertação. Verifica-se que a abordagem utilizando RNA tipo MLP obteve melhor acurácia geral em comparação com as outras três, com 71,08%, seguido pela abordagem usando RNA tipo RBF com 64,32% e pela abordagem utilizando Sistema *Fuzzy* com as adaptações. A pior acurácia foi obtida pela abordagem TOP(X) original, com apenas 52,7%.

Tabela 19 – Precisão total de cada abordagem

Abordagem	Acurácia
Original	52,7%
Sistema *Fuzzy* com adaptações	60,54%
RNA tipo MLP	**71,08%**
RNA tipo RBF	64,32%

Entretanto, avaliar a performance do modelo de classificação apenas com a medida de acurácia não é plenamente aceito pela comunidade, pois ela é considerada uma medida fraca (PROVOST; FAWCETT, 1997), além de que, um modelo cujo objetivo é maximizar a acurácia pode aparentemente ter uma boa avaliação, pois pode considerar as informações

irrelevantes (MANNING; RAGHAVAN; SCHüTZE, 2008). Por este motivo, resolveu-se analisar as abordagens por meio de suas classes, ou seja, pelos seus graus de importância com outras medidas de avaliação. A primeira medida analisada é a medida-F, mostrada na Tabela 20.

Tabela 20 – Medida-F por classe de cada abordagem

Classe	Original	Sistema *Fuzzy* com adaptações	RNA tipo MLP	RNA tipo RBF
EXC	35,95%	**76,41%**	71,4%	73%
BM	24,82%	55,3%	**69,7%**	65,6%
SF	48,53%	47,4%	**63,2%**	50,4%
ISF	77,15%	71,3%	**79,8%**	73,9%

Pode-se observar que com relação aos comentários da classe EXC (excelente), a abordagem com Sistema *Fuzzy* e adaptações obteve melhor valor de medida-F com 76,41%, enquanto a abordagem com RNA tipo MLP obteve melhores valores de medida-F nas classes restantes. É importante ressaltar que a abordagem TOP(X) original obteve os piores resultados em todas as classes, exceto na classe ISF, referente aos comentários insuficientes, no qual obteve medida-F de 77,15%, apenas atrás da abordagem com RNA tipo MLP.

Uma outra medida de avaliação utilizada foi o Coeficiente de Correlação de Matthews. Proposta por Matthews (1975), é uma medida que leva em conta os verdadeiros e falsos positivos e é considerada uma medida equilibrada que pode ser utilizada mesmo se as classes são de tamanhos diferentes. O cálculo do Coeficiente de Correlação de Matthews retorna um valor entre -1 e +1, onde o coeficiente de +1 representa uma predição perfeita, 0 representa uma predição aleatória média e -1 uma predição inversa. A equação que retorna o coeficiente é

$$\varphi = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.4}$$

onde o valor TN representa os comentários corretamente classificados como não sendo da classe analisada (do inglês: *true negative*). É a medida que é mais precisa para a comparação das abordagens propostas nesta Dissertação, sendo geralmente considerada como sendo uma das melhores métricas para se medir a performance de um sistema (POWERS, 2011), além de o valor desta métrica possuir uma interpretação que indica o quão próximo da perfeição o algoritmo se encontra. Os valores referentes a cada classe são apresentados na Tabela 21.

Assim como verificado nos resultados referentes à medida-F, os coeficientes de correlação por cada classe de cada abordagem mostra que o Sistema *Fuzzy* com adaptações

Tabela 21 – Coeficiente de Correlação de Matthews por classe de cada abordagem

Classe	Original	Sistema *Fuzzy* com adaptações	RNA tipo MLP	RNA tipo RBF
EXC	0,38	**0,77**	0,70	0,72
BM	0,10	0,46	**0,59**	0,53
SF	0,24	0,23	**0,44**	0,28
ISF	0,64	0,54	**0,69**	0,59

obteve o melhor valor em relação às outras abordagens no que se referencia aos comentários da classe EXC. A abordagem com RNA tipo MLP alcançou melhores coeficientes nas classes restantes. Mais uma vez a abordagem original obteve piores coeficientes na maioria das classes.

Por fim, uma análise de performance das abordagens a partir da orientação semântica dos comentários foi realizada. A Tabela 22 mostra a medida-F de cada classe, divididas pela polaridade do comentário: positivo, negativo ou neutro. Ressalta-se que nessa análise foram unidos os comentários considerados excelentes e bons em um única classe pelo fato de essas classes terem um pequeno número de comentários para uma análise individual, sendo denominada Excelentes e Bons (EXC+BM).

Tabela 22 – Medida-F por polaridade de cada classe de cada abordagem

Classe	Original	Sistema *Fuzzy* com adaptações	RNA tipo MLP	RNA tipo RBF
POSITIVOS				
EXC+BM	25,2%	58,82%	**71,60%**	62,5%
SF	50%	41,37%	**60%**	50%
ISF	73,6%	72,97%	**81,35%**	76,4%
NEGATIVOS				
EXC+BM	25,02%	50%	**62,68%**	50%
SF	52,17%	42,5%	**70,21%**	52,74%
ISF	84,74%	64,1%	**86,23%**	79,27%
NEUTROS				
EXC+BM	25,9%	**76,36%**	74,19%	61,76%
SF	41,26%	52,63%	**62,5%**	37,4%
ISF	68,49%	72,97%	**80%**	74,28%

Verifica-se que em todos as classes de todas as polaridades a abordagem utilizando o modelo computacional baseado em RNA tipo MLP se comportou melhor, exceto na detecção dos melhores comentários na polaridade neutra. Ao focar mais detalhadamente

na classe EXC+BM, a classe onde se encontram os comentários mais importantes, a abordagem RNA tipo MLP tem números de medida-F superiores às das outras abordagens nas polaridades positiva e negativa, o que permite interpretar que a abordagem em questão é mais precisa quando há a definição de polaridade do comentário.

A partir da análise das três medidas calculadas neste experimento, a comparação das abordagens permite tirar as seguintes interpretações:

- As duas abordagens utilizando RNA obtiveram melhores resultados na maioria das medidas, independente das classes;
- Para análise dos comentários **mais importantes**, ou seja, os comentários considerados excelentes, a abordagem com **Sistema *Fuzzy*** e adaptações obteve melhores resultados;
- Para análise dos comentários **positivos ou negativos mais importantes**, a abordagem com **RNA MLP** obteve melhores resultados;

Portanto, em comparação de medidas estatísticas entre as abordagens propostas nesta Dissertação, com todas as adaptações e justificativas realizadas nos capítulos anteriores, pode-se observar que a abordagem que melhor identifica os comentários mais importantes de um Córpus de comentários sobre hotéis é utilizando Sistema *Fuzzy*, enquanto a abordagem que melhor identifica os comentários mais importantes de um Córpus de comentários sobre hotéis anotado com sua orientação semântica (positivo ou negativo) é utilizando Redes Neurais Artificiais MLP.

4.4 Considerações Finais

Este Capítulo apresentou o experimento das abordagens com Sistema *Fuzzy* e Redes Neurais Artificiais tipo MLP e RBF propostas nesta Dissertação a partir das adaptações realizadas nas variáveis de entrada referentes à quantidade de tuplas e riqueza do vocabulário, além da utilização da abordagem TOP(X) original proposta por Sousa (2015) e fez a discussão dos resultados a fim de comparar ambas as abordagens.

Análises estatísticas usando matrizes de confusão foram extraídas a fim de comparar as abordagens. Uma análise de importância das variáveis de entrada em relação à saída das abordagens utilizando RNA foi realizada, obtendo as variáveis referentes à quantidade de tuplas e a reputação do autor como as mais importantes na estimação da importância de um comentário.

Na comparação das abordagens por meio dos métodos de avaliação de precisão total (acurácia), medida-F e Coeficiente de Correlação de Matthews, foi observado que a abordagem com Sistema *Fuzzy* com as adaptações propostas obteve melhores resultados na

detecção dos comentários considerados excelentes pelo subcórpus de importância, enquanto a abordagem com RNA tipo MLP foram mais eficientes ao detectar comentários mais importantes quando houve a definição de polaridade do comentário..

Conclusões

De forma a automatizar a tarefa de analisar descrições textuais devido à grande quantidade de comentários na *Web*, este trabalho teve como objetivo central comparar abordagens para resolver o problema de estimação do grau de importância utilizando modelos computacionais baseados em Sistema *Fuzzy* e Redes Neurais Artificiais (RNA), tipos MLP e RBF. Tais modelos computacionais são paramétricos, o que fez com que esta comparação entre eles fosse cuidadosamente planejada e realizada.

A abordagem base para este trabalho foi proposta por Sousa (2015). Objetivos específicos desta Dissertação tinham como propósito adaptar duas das variáveis de entrada da abordagem base: quantidade de tuplas <característica, palavra opinativa> e riqueza do vocabulário.

A utilização de uma base referente às características de hotéis e melhorias nos padrões linguísticos foram os pontos chaves referentes à variável de entrada quantidade de tuplas. Referente à variável de entrada riqueza do vocabulário, três novas métricas foram adicionadas: o índice TTR, índice Maas e o índice MTLD. Esses índices são importantes para mensurar a riqueza do vocabulário de textos, ponderando termos pelas palavras existentes no comentário. Sem esquecer da corretude do comentário, a adaptação proposta passou pela implementação de uma RNA na qual as entradas seriam os valores da corretude do comentário, índice TTR, índice Maas e o índice MTLD, na qual a saída mede o grau de riqueza do vocabulário do comentário, servindo como variável de entrada para a abordagem usando Sistema *Fuzzy*. As quatro variáveis também serviram como entrada para a abordagem utilizando RNA.

Vale ressaltar que a variável de entrada reputação do autor não foi explorada neste trabalho. A utilização dessa variável é feita pelo número de comentários realizados no site *Booking.com*. Destaca-se que o tratamento desta variável está sendo explorado por um aluno do grupo de pesquisa que encontra-se atualmente cursando mestrado no PPGCC/UFPI.

Além das adaptações das variáveis de entrada, a abordagem utilizando Sistema *Fuzzy* foi modificada para suprir as adaptações das variáveis de entrada. A base de regras e as funções de pertinência foram ajustadas. Vale a ressalva que um dos objetivos das adaptações nas variáveis era exatamente melhorar a performance do sistema proposto por Sousa (2015), o qual foi alcançado.

Em relação à abordagem utilizando RNA, dois tipos foram propostos por resolverem o problema de classificação de padrões: MLP e RBF. Em todos os graus de importância, a RNA MLP se comportou melhor com relação às medidas de precisão, cobertura e medida-

F. Além disso, foram analisadas também quais das variáveis de entrada da RNA MLP eram consideradas mais importantes para definir o grau de importância dos comentários. Quantidade de tuplas e reputação do autor foram as consideradas mais importantes em testes realizados utilizando técnicas de seleção de características.

Por fim, foram comparadas as abordagens utilizando os modelos computacionais inspirados em Sistemas *Fuzzy* e Redes Neurais Artificiais MLP após o experimento. As matrizes de confusão geradas pelas execuções de ambas as abordagens foram utilizadas para realizar uma comparação estatística de acurácia, precisão, cobertura, medida-F, eficiência e correlação. Interpretou-se que nas análises dos comentários mais importantes (comentários excelentes) a abordagem utilizando Sistema *Fuzzy* obteve melhor performance, enquanto para detectar os comentários mais importantes quando definida a orientação semântica de cada comentário (positivo ou negativo), a abordagem utilizando RNA MLP teve melhores resultados.

Algumas limitações podem ser citadas, a saber: i) a dificuldade de construir um Córpus anotado passa pela falta de padrão na escolha das classes dos comentários (excelentes e bons, por exemplo) pelos especialistas. Como alternativas, a utilização de exemplos de classificação tais como um conjunto de regras para classificar o comentário em tal classe e o uso da Escala Likert (LIKERT, 1932); e ii) o desempenho das ferramentas utilizadas no trabalho, como o etiquetador e os padrões linguísticos. O primeiro pode ser visto em relação às características que não foram corretamente devido a erros de escrita e palavras fora do contexto formal, tal como o internetês[1] ou erros de grafia. O segundo pode ser explicado pelo fato da detecção manual de tais padrões linguísticos, nos quais nem todos os possíveis estão listados, o que poderia ser feito automaticamente por meio de métodos existentes na literatura.

Trabalhos Futuros

Alguns trabalhos futuros são propostos após o resultado desta pesquisa. Eis:

- Executar as abordagens utilizando RNAs excluindo as variáveis menos importantes mostradas na análise dos modelos presentes nos resultados desta Dissertação;
- Realizar um estudo mais profundo na variável de Reputação do Autor e aplicá-las na abordagem proposta nesta Dissertação com as adaptações;
- Definir outras métricas para o cálculo do grau de importância do comentário, como o tempo que o comentário foi realizado e o tamanho do texto;

[1] O internetês é conhecido como forma grafolinguística que se difundiu em textos como *chats*, *blogs* e demais RSOs como uma prática de escrita caracterizada pelo registro divergente da norma culta padrão, tomado como "simplificação da escrita". (KOMESU; TENANI, 2009)

- Propor uma nova abordagem utilizando Sistemas Neuro-Fuzzy, a qual se integram os dois modelos computacionais explorados nesta Dissertação, além de outras técnicas de aprendizagem de máquina;
- Estudar os erros de detecção nos comentários, os quais podem ser resolvidos por meio de construção de bases de gírias, ironias e sarcasmos, baseados em normalização de textos (AVANÇO, 2015);
- Construir uma ferramenta (aplicação) que receba comentários do usuário e retorne os comentários mais importantes, de acordo com a escolha do usuário;

Finalmente, é importante ressaltar que este trabalho faz parte de um projeto maior que objetiva analisar informações de produtos e/ou serviços tomando base de três fontes: *sites* de fabricantes de produtos e/ou serviços, *sites* de vendas e o *sites* de reclamações. Dessa forma, busca-se gerar um conhecimento mais amplo ao avaliar o produto oferecido por uma empresa, por meio da comparação entre opiniões comuns e opiniões notadamente negativas. Destaca-se também que outros alunos de mestrado do PPGCC/UFPI já estão realizando pesquisas nas dimensões de reputação do autor e em novos modelos computacionais, como as Redes Neurais Convolucionais (LECUN; BENGIO, 1998).

Referências

ACIAR, S.; ZHANG, D.; SIMOFF, S.; DEBENHAM, J. Informed recommender: Basing recommendations on consumer product reviews. *IEEE Intelligent Systems*, v. 22, n. 3, p. 39–47, 2007. Citado na página 33.

ALUÍSIO, S.; PELIZONNI, J.; MARCHI, A.; OLIVEIRA, L.; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: *6th Int. Conf. on Computacional Processing Of Portuguese Language (PROPOR)*. [S.l.: s.n.], 2003. p. 110–117. Citado na página 24.

ARCHAK, N.; GHOSE, A.; IPEIROTIS, P. G. Show me the money!: Deriving the pricing power of product features by mining consumer reviews. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2007. p. 56–65. Citado na página 3.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. [S.l.: s.n.], 2010. p. 492–499. Citado na página 3.

AVANÇO, L. V. *Sobre normalização e classificação de polaridade de textos opinativos na web*. 102 p. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação (ICMC/USP), 2015. Citado na página 67.

BAI, Y.; WANG, D. Fundamentals of fuzzy logic control—fuzzy sets, fuzzy rules and defuzzifications. In: *Advanced Fuzzy Logic Technologies in Industrial Applications*. [S.l.]: Springer, 2006. p. 17–36. Citado na página 16.

BALAHUR, A.; STEINBERGER, R.; GOOT, E. v. d.; POULIQUEN, B.; KABADJOV, M. Opinion mining on newspaper quotations. In: *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. [S.l.: s.n.], 2009. p. 523–526. Citado na página 3.

BARBOSA, J. L.; MOURA, R. S.; SANTOS, R. L. d. S. Predicting portuguese steam review helpfulness using artificial neural networks. In: *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*. [S.l.: s.n.], 2016. p. 287–293. Citado na página 2.

BECKER, K.; TUMITAN, D. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In: *Anais do 28 Simpósio Brasileiro de Banco de Dados*. [S.l.: s.n.], 2013. Citado 3 vezes nas páginas 19, 12 e 13.

BERRY, M. J.; LINOFF, G. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. [S.l.]: John Wiley & Sons, Inc., 1997. Citado na página 3.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly, 2012. Citado 4 vezes nas páginas 21, 9, 23 e 24.

BOLLEGALA, D.; MATSUO, Y.; ISHIZUKA, M. A web search engine-based approach to measure semantic similarity between words. *IEEE Transactions on Knowledge and Data Engineering*, v. 23, n. 7, p. 977–990, 2011. Citado na página 1.

BONCHI, F.; CASTILLO, C.; GIONIS, A.; JAIMES, A. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology*, v. 2, n. 3, 2011. Citado na página 2.

BRAGA, A. d. P.; CARVALHO, A. P. d. L. F. d.; LUDERMIR, T. B. *Redes neurais artificiais: teoria e aplicações*. [S.l.]: LTC Editora, 2007. Citado na página 47.

BRIDLE, J. S. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: *Advances in Neural Information Processing Systems 2*. [S.l.]: Morgan Kaufmann Publishers Inc., 1990. p. 211–217. Citado na página 19.

BROOMHEAD, D. S.; LOWE, D. *Radial basis functions, multi-variable functional interpolation and adaptive networks*. [S.l.], 1988. Citado na página 21.

BRUCE, R. F.; WIEBE, J. M. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, v. 5, n. 2, p. 187–205, 1999. Citado na página 33.

BUGMANN, G. Normalized gaussian radial basis function networks. *Neurocomputing*, v. 20, n. 1, p. 97–110, 1998. Citado na página 49.

CARROLL, J. B. *Language and thought*. [S.l.]: Prentice-Hall Englewood Cliffs, 1964. Citado na página 6.

CHAMLERTWAT, W.; BHATTARAKOSOL, P.; RUNGKASIRI, T.; HARUECHAIYASAK, C. Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, v. 18, n. 8, p. 973–992, 2012. Citado na página 1.

CHEN, G.; PHAM, T. *Introduction to Fuzzy Systems*. [S.l.]: CRC Press, 2005. Citado na página 17.

CHEN, S.; COWAN, C. F.; GRANT, P. M. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, v. 2, n. 2, p. 302–309, 1991. Citado na página 47.

CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, v. 78, p. 124–134", 2017. Citado na página 36.

COX, E. *The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems*. [S.l.]: AP Professional, 1994. Citado na página 17.

DASGUPTA, D. Advances in artificial immune systems. *IEEE Computational Intelligence Magazine*, v. 1, n. 4, p. 40–49, 2006. Citado na página 15.

DAVE, K.; LAWRENCE, S.; ; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: *Proceedings of the 12th International Conference on World Wide Web*. [S.l.: s.n.], 2003. p. 519–528. Citado na página 3.

DIAS-DA-SILVA, B. C. Brazilian portuguese wordnet: A computational linguistic exercise of encoding bilingual relational lexicons. *International Journal of Computational Linguistics and Applications*, v. 1, n. 1-2, p. 137–150, 2010. Citado na página 33.

DING, X.; LIU, B.; YU, P. S. A holistic lexicon-based approach to opinion mining. In: *Proceedings of the International Conference on Web Search and Data Mining*. [S.l.: s.n.], 2008. p. 231–240. Citado na página 32.

D'INVERNO, M.; LUCK, M. *Understanding Agent Systems*. [S.l.]: Springer, 2004. Citado na página 15.

DRAGONI, M.; TETTAMANZI, A. G. B.; PEREIRA, C. D. C. A fuzzy system for concept-level sentiment analysis. In: *Semantic Web Evaluation Challenge*. [S.l.: s.n.], 2014. v. 475, p. 21–27. Citado na página 33.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 2014. Citado na página 49.

DUCH, W.; JANKOWSKI, N. Survey of neural transfer functions. *Neural Computing Surveys*, v. 2, n. 1, p. 163–212, 1999. Citado na página 49.

ESULI, A.; SEBASTIANI, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In: *Proceedings of the 5th Conference on Language Resources and Evaluation*. [S.l.: s.n.], 2006. p. 417–422. Citado na página 35.

FELDMAN, R. Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM*, v. 56, n. 4, p. 82–89, 2013. Citado na página 1.

FELIPPO, A. D.; TOSTA, F.; PARDO, T. Applying lexical-conceptual knowledge for multilingual multi-document summarization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 9727, p. 38–49, 2016. Citado na página 1.

FELLBAUM, C. *WordNet: An Eletronic Lexical Database*. [S.l.]: The MIT Press, 1998. Citado na página 33.

FU, G.; WANG, X. Chinese sentence-level sentiment classification based on fuzzy sets. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. [S.l.: s.n.], 2010. p. 312–319. Citado na página 34.

GHANI, R.; PROBST, K.; LIU, Y.; KREMA, M.; FANO, A. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, v. 1, n. 8, p. 41–48, 2006. Citado na página 3.

GIL DE ZÚÑIGA, H. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, v. 17, n. 3, p. 319–336, 2012. Citado na página 2.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. Large-scale sentiment analysis for news and blogs. In: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. [S.l.: s.n.], 2007. Citado na página 3.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. ed. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1989. Citado na página 15.

GOLDENBERG, J.; LIBAI, B.; MULLER, E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, v. 12, n. 3, p. 211–223, 2001. Citado na página 2.

GUERRA, P. H. C.; VELOSO, A.; JR, W. M.; ALMEIDA, V. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2011. p. 150–158. Citado na página 3.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, v. 46, n. 1-3, p. 389–422, 2002. Citado na página 58.

HAMBURG, M. *Basic Statistics: A Modern Approach*. [S.l.]: Harcourt Brace Jovanovich, 1985. Citado na página 27.

HATZIVASSILOGLOU, V.; WIEBE, J. Effects of adjective orientation and gradability on sentence subjectivity. In: *Proceedings of 18th International Conference on Computational Linguistics*. [S.l.: s.n.], 2000. p. 299–305. Citado na página 31.

HAYKIN, S. *Neural Networks and Learning Machines*. [S.l.]: Prentice Hall, 2009. Citado 2 vezes nas páginas 4 e 22.

HE, X.; LAPEDES, A. Nonlinear modeling and prediction by successive approximation using radial basis functions. *Physica D: Nonlinear Phenomena*, v. 70, n. 3, p. 289 – 301, 1994. Citado na página 47.

HEBB, D. O. *The organization of behavior: A neuropsychological theory*. [S.l.]: Wiley, 1949. Citado na página 21.

HEIMES, F.; HEUVELN, B. van. The normalized radial basis function neural network. In: *IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.: s.n.], 1998. v. 2, p. 1609–1614. Citado na página 49.

HODGKIN, A. L.; HUXLEY, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, v. 117, n. 1, p. 500–544, 1952. Citado na página 18.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, v. 79, n. 8, p. 2554–2558, 1982. Citado na página 21.

HU, M.; LIU, B. Mining opinion features in customer reviews. In: *Proceedings of the 19th National Conference on Artifical Intelligence*. [S.l.: s.n.], 2004. p. 755–760. Citado na página 3.

IBM Corp. *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, NY: IBM Corp., 2011. Citado na página 43.

INDHUJA, K.; REGHU, R. P. C. Fuzzy logic based sentiment analysis of product review documents. In: *First International Conference on Computational Systems and Communications (ICCSC)*. [S.l.: s.n.], 2014. p. 18–22. Citado na página 33.

JACKSON, P.; MOULINIER, I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. [S.l.]: John Benjamins Publishing Company, 2002. Citado na página 9.

JANG, J.-S. R.; SUN, C.-T. *Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. [S.l.]: Prentice-Hall, Inc., 1997. Citado na página 17.

JEONG, H.; SHIN, D.; CHOI, J. Ferom: Feature extraction and refinement for opinion mining. *Electronics and Telecommunications Research Institute Journal*, v. 33, n. 5, p. 720–730, 2011. Citado na página 32.

JINDAL, N.; LIU, B. Identifying comparative sentences in text documents. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.: s.n.], 2006. p. 244–251. Citado na página 10.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. In: WILLETT, P. (Ed.). *Document Retrieval Systems*. [S.l.: s.n.], 1988. p. 132–142. Citado na página 33.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 2nd. ed. [S.l.]: Pearson Prentice Hall, 2008. Citado 2 vezes nas páginas 1 e 9.

KAMPS, J.; MARX, M.; MOKKEN, R. J.; RIJKE, M. de. Using wordnet to measure semantic orientations of adjectives. In: *In Proceedings of 4th international conference on language resources and evaluation*. [S.l.: s.n.], 2004. p. 1115–1118. Citado na página 33.

KAR, A.; MANDAL, D. P. Finding opinion strength using fuzzy logic on web reviews. *International Journal of Engineering and Industries*, v. 2, n. 1, p. 37–43, 2011. Citado na página 34.

KASPER, W.; VELA, M. Sentiment analysis for hotel reviews. In: *Proccedings of the Computational Linguistics-Applications Conference*. [S.l.: s.n.], 2011. p. 45–52. Citado na página 3.

KENNEDY, J.; EBERHART, R. C. *Swarm Intelligence*. [S.l.]: Morgan Kaufmann Publishers Inc., 2001. Citado na página 15.

KHAN, K.; BAHARUDIN, B.; KHAN, A.; MALIK, F. e. Mining opinion from text documents: A survey. In: *3rd IEEE International Conference on Digital Ecosystems and Technologies*. [S.l.: s.n.], 2009. p. 217–222. Citado na página 36.

KIM, Y. A.; SRIVASTAVA, J. Impact of social influence in e-commerce decision making. In: *Proceedings of the Ninth International Conference on Electronic Commerce*. [S.l.: s.n.], 2007. p. 293–302. Citado na página 2.

KLAVANS, J.; KAN, M.-Y. Role of verbs in document analysis. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. [S.l.: s.n.], 1998. p. 680–686. Citado na página 40.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 1995. p. 1137–1143. Citado na página 48.

KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, v. 43, n. 1, p. 59–69, 1982. Citado na página 21.

KOLEKAR, N.; RAO, G.; DEY, S.; MANE, M.; JADHAV, V.; PATIL, S. Sentiment analysis and classification using lexicon-based approach and addressing polarity shift problem. *Journal of Theoretical and Applied Information Technology*, v. 90, n. 1, p. 118–125, 2016. Citado na página 35.

KOMESU, F.; TENANI, L. Considerações sobre o conceito de "internetês"nos estudos da linguagem. *Linguagem em (Dis)curso*, v. 9, p. 621 – 643, 2009. Citado na página 66.

KRIESEL, D. *A Brief Introduction to Neural Networks*. [s.n.], 2007. Disponível em: <http://www.dkriesel.com>. Citado 3 vezes nas páginas 18, 22 e 47.

KU, L.; LIANG, Y.; CHEN, H. Opinion extraction, summarization and tracking in news and blog corpora. In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. [S.l.: s.n.], 2006. p. 100–107. Citado na página 3.

KUCUKTUNC, O.; CAMBAZOGLU, B.; WEBER, I.; FERHATOSMANOGLU, H. A large-scale sentiment analysis for yahoo! answers. In: *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. [S.l.: s.n.], 2012. p. 633–642. Citado na página 1.

KURIAN, N.; ASOKAN, S. Summarizing user opinions: A method for labeled-data scarce product domains. *Procedia Computer Science*, v. 46, p. 93–100, 2015. Citado na página 35.

KWOK, C.; ETZIONI, O.; WELD, D. Scaling question answering to the web. *ACM Transactions on Information Systems*, v. 19, n. 3, p. 242–262, 2001. Citado na página 1.

LECUN, Y.; BENGIO, Y. Convolutional networks for images, speech, and time series. In: *The Handbook of Brain Theory and Neural Networks*. [S.l.: s.n.], 1998. p. 255–258. Citado na página 67.

LI, M.; HUANG, L.; TAN, C.-h.; WEI, K. K. Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, v. 17, n. 4, p. 101–136, 2013. Citado na página 4.

LIKERT, R. A technique for the measurement of attitudes. *Archives of Psychology*, v. 22, n. 140, p. 1–55, 1932. Citado na página 66.

LIU, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Secaucus, USA: Springer-Verlag New York, Inc., 2006. Citado na página 12.

LIU, B. Sentiment analysis and subjectivity. In: *Handbook of natural language processing*. [S.l.]: Chapman and Hall/CRC, 2010. Citado 4 vezes nas páginas 1, 3, 10 e 11.

LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan and Claypool, 2012. Citado 3 vezes nas páginas 11, 13 e 14.

LIU, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. [S.l.]: Cambridge University Press, 2015. Citado na página 14.

LIU, H.; SINGH, P. Conceptnet: A practical commonsense reasoning tool-kit. *BT Technology Journal*, v. 22, n. 4, p. 211–226, 2004. Citado na página 38.

LIU, J.; WU, G.; YAO, J. Opinion searching in multi-product reviews. In: *6th IEEE International Conference on Computer and Information Technology*. [S.l.: s.n.], 2006. p. 25–26. Citado na página 31.

LOUSADA, M.; VALENTIM, M. L. P. Modelos de tomada de decisão e sua relação com a informação orgânica. *Perspectivas em Ciência da Informação*, v. 16, p. 147–164, 2011. Citado na página 1.

LU, X. *Computational Methods for Corpus Annotation and Analysis*. [S.l.]: Springer Netherlands, 2014. Citado na página 42.

LUGER, G. F. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. 4th. ed. [S.l.]: Addison-Wesley Publishing Company, 2004. Citado na página 14.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, v. 1, n. 4, p. 309–317, 1957. Citado na página 33.

MAGOULAS, G. D.; VRAHATIS, M. N.; ANDROULAKIS, G. S. Improving the convergence of the backpropagation algorithm using learning rate adaptation methods. *Neural Computation*, v. 11, n. 7, p. 1769–1796, 1999. Citado na página 43.

MAK, M.; ALLEN, W.; SEXTON, G. Comparing multi-layer perceptrons and radial basis functions networks in speaker recognition. *Journal of Microcomputer Applications*, v. 16, n. 2, p. 147–159, 1993. Citado na página 47.

MAMDANI, E.; ASSILIAN, S. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, v. 7, n. 1, p. 1–13, 1975. Citado 2 vezes nas páginas 17 e 44.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008. Citado na página 60.

MASS, H. D. Zusammenhang zwischen wortschatzumfang und lange eines textes. In: *Zeitschrift fur Literaturwissenschaft und Linguistik*. [S.l.: s.n.], 1972. p. 73–79. Citado na página 41.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, v. 405, n. 2, p. 442–451, 1975. Citado na página 60.

MCCARTHY, P. M. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Tese (Doutorado) — The University of Memphis, 1993. Citado na página 41.

MCCARTHY, P. M.; JARVIS, S. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, v. 42, n. 2, p. 381–392, 2010. Citado 2 vezes nas páginas 41 e 42.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, n. 4, p. 115–133, 1943. Citado na página 18.

MCNEILL, M.; RAESIDE, R.; GRAHAM, M.; ROSEBOOM, I. Comparing summarisation techniques for informal online reviews. In: *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. [S.l.: s.n.], 2015. p. 322–329. Citado na página 36.

MICHALEWICZ, Z. *Genetic Algorithms + Data Structures = Evolution Programs*. [S.l.]: Springer-Verlag, 1996. Citado na página 15.

MILNE, D.; WITTEN, I. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, v. 194, p. 222–239, 2013. Citado na página 2.

MITCHELL, T. M. *Machine Learning*. 1st. ed. [S.l.]: McGraw-Hill, Inc., 1997. Citado na página 14.

MULLEN, T.; COLLIER, N. Sentiment analysis using support vector machines with diverse information sources. In: *Proceedings of the Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2004. p. 412–418. Citado na página 33.

NADALI, S.; MURAD, M. A. A.; KADIR, R. A. Sentiment classification of customer reviews based on fuzzy logic. In: *International Symposium on Information Technology*. [S.l.: s.n.], 2010. v. 2, p. 1037–1044. Citado na página 34.

NAVIGLI, R.; PONZETTO, S. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, v. 193, p. 217–250, 2012. Citado na página 1.

NEDJAH, N.; MOURELLE, L. de M. *Fuzzy Systems Engineering: Theory and Practice*. [S.l.]: Springer Berlin Heidelberg, 2010. Citado na página 17.

OLIVEIRA, P.; SOUZA, M.; BRAGA, R.; BRITTO, R.; RABELO, R. L.; NETO, P. Athena: A visual tool to support the development of computational intelligence systems. In: *IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI)*. [S.l.: s.n.], 2014. p. 950–959. Citado na página 24.

OLIVEIRA, P. A. M. de. *Athena: uma Arquitetura e uma Ferramenta para Auxiliar o Desenvolvimento de Sistemas Baseados em Inteligência Computacional*. 68-70 p. Dissertação (Mestrado) — Universidade Federal do Piauí, 2016. Citado na página 25.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. [S.l.: s.n.], 2002. p. 79–86. Citado na página 33.

PARKS, R.; LEVINE, D.; LONG, D. *Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience*. [S.l.]: MIT Press, 1998. Citado na página 22.

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, v. 50, n. 302, p. 157–175, 1900. Citado na página 58.

PEDRYCZ, W.; CHEN, S. *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*. [S.l.]: Springer International Publishing, 2016. Citado na página 14.

PEDRYCZ, W.; GOMIDE, F. *Fuzzy Systems Engineering: Toward Human-Centric Computing*. [S.l.]: Wiley-IEEE Press, 2007. Citado na página 15.

PIANOSI, F.; BEVEN, K.; FREER, J.; HALL, J. W.; ROUGIER, J.; STEPHENSON, D. B.; WAGENER, T. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, v. 79, p. 214–232, 2016. Citado na página 58.

PIATETSKY-SHAPIRO, G. Knowledge discovery in real databases: A report on the IJCAI-89 workshop. *AI Magazine*, v. 11, n. 5, p. 68–70, 1991. Citado na página 3.

PISKORSKI, J.; YANGARBER, R. Information extraction: Past, present and future. In: *Multi-source, Multilingual Information Extraction and Summarization*. [S.l.]: Springer Berlin Heidelberg, 2013. p. 23–49. Citado na página 54.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Citado 2 vezes nas páginas 53 e 60.

POZZI, F.; FERSINI, E.; MESSINA, E.; LIU, B. *Sentiment Analysis in Social Networks*. [S.l.]: Elsevier Science & Technology Books, 2016. Citado na página 14.

PROVOST, F.; FAWCETT, T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1997. p. 43–48. Citado na página 59.

RAUT, V.; LONDHE, D. Opinion mining and summarization of hotel reviews. In: *Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks*. [S.l.: s.n.], 2014. p. 556–559. Citado na página 36.

RAUT, V.; LONDHE, D. Survey on opinion mining and summarization of user reviews on web. *International Journal of Computer Science and Information Technologies*, v. 5, n. 2, p. 1026–1030, 2014. Citado na página 36.

RIJSBERGEN, C. J. V. *Information Retrieval*. 2nd. ed. Newton: Butterworth-Heinemann, 1979. Citado na página 54.

ROBERTSON, D. W. A note on the classical origin of "circumstances"in the medieval confessional. *Studies in Philology*, v. 43, n. 1, p. 6–14, 1946. Citado na página 9.

ROJAS, R. *Neural Networks: A Systematic Introduction*. [S.l.]: Springer-Verlag New York, Inc., 1996. Citado na página 17.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, p. 65–386, 1958. Citado na página 21.

ROSS, T. *Fuzzy Logic with Engineering Applications*. 3rd. ed. [S.l.]: Wiley, 2004. Citado 2 vezes nas páginas 14 e 15.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. [S.l.: s.n.], 1986. p. 318–362. Citado na página 34.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Pearson Education, 2009. Citado na página 59.

SAMARASINGHE, S. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*. [S.l.]: CRC Press, 2016. Citado na página 22.

SANTOS, G. C. dos; THOMAZ, P. S.; RIBEIRO, F. M.; ARAÚJO, J. F.; MATTOS, V. L. D. de. Influência do método de defuzzificação em mensurações com controladores fuzzy. *Blucher Marine Engineering Proceedings*, v. 1, n. 1, p. 845–852, 2014. Citado na página 46.

SANTOS, R. L. de S.; MOURA, R. S. Extração de métricas e análise de sentimentos em comentários web no domínio de hotéis. In: *Proceedings of the 5th Brazilian Workshop on Social Network Analysis and Mining*. [S.l.: s.n.], 2016. p. 127–138. Citado na página 6.

SANTOS, R. L. de S.; SOUSA, R. F. de; RABELO, R. A. L.; MOURA, R. S. An experimental study based on fuzzy systems and artificial neural networks to estimate the importance of reviews about product and services. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2016. p. 647–653. Citado 3 vezes nas páginas 6, 14 e 34.

SANTOS, R. L. de S.; VIEIRA, J. ao P. A.; BARBOSA, J. L. N.; SÁ, C. A.; MOURA, E. G.; MOURA, R. S.; SOUSA, R. F. de. Evaluating the importance of web comments through metrics extraction and opinion mining. In: *Proceedings of the 35th International Conference of the Chilean Computer Science Society*. [S.l.: s.n.], 2016. p. 153–163. Citado na página 6.

SEERAT, B.; AZAM, F. Opinion mining: Issues and challenges (a survey). *International Journal of Computer Applications*, v. 49, n. 9, p. 42–51, 2012. Citado na página 36.

SERGIENKO, R.; AKHTIAMOV, O.; SEMENKIN, E.; SCHMITT, A. A novel approach to neural network design for natural language call routing. In: *Proccedings of 12th International Conference on Informatics in Control, Automation and Robotics*. [S.l.: s.n.], 2015. v. 1, p. 102–109. Citado na página 14.

SHARMA, A.; DEY, S. An artificial neural network based approach for sentiment analysis of opinionated text. In: *Proceedings of the 2012 ACM Research in Applied Computation Symposium*. [S.l.: s.n.], 2012. p. 37–42. Citado na página 33.

SHARMA, A.; DEY, S. A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. *SIGAPP Applied Computing Review*, v. 12, n. 4, p. 67–75, 2012. Citado na página 33.

SHI, H. X.; LI, X. J. A sentiment analysis model for hotel reviews based on supervised learning. In: *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*. [S.l.: s.n.], 2011. v. 3, p. 950–954. Citado na página 32.

SILER, W.; BUCKLEY, J. *Fuzzy Expert Systems and Fuzzy Reasoning*. [S.l.]: Wiley, 2005. Citado 2 vezes nas páginas 14 e 15.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. *Redes Neurais Artificiais: para engenharia e ciencias aplicadas*. [S.l.]: Artliber, 2010. Citado 6 vezes nas páginas 19, 17, 18, 20, 22 e 47.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*. [S.l.: s.n.], 2012. p. 218–228. Citado na página 33.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: *4th International Workshop on Web and Text Intelligence*. [S.l.: s.n.], 2012. Citado 2 vezes nas páginas 32 e 36.

SIMON, H. A. Why should machines learn? In: *Machine Learning: An Artificial Intelligence Approach*. [S.l.]: Springer Berlin Heidelberg, 1983. p. 25–37. Citado na página 14.

SOUSA, R. F. de. *Abordagem TOP(X) para Inferir Comentários mais Importantes sobre Produtos e Serviços*. Dissertação (Mestrado) — Universidade Federal do Piauí, 2015. Citado 16 vezes nas páginas 15, 17, 19, 4, 6, 33, 34, 36, 37, 38, 41, 42, 44, 45, 62 e 65.

SOUZA, M.; VIEIRA, R.; BUSETTI, D.; CHISHMAN, R.; ALVES, I. M. Construction of a portuguese opinion lexicon from multiple resources. In: *In 8th Brazilian Symposium in Information and Human Language Technology - STIL*. [S.l.: s.n.], 2011. Citado na página 33.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. [S.l.]: MIT Press, 1998. Citado na página 21.

TAN, C.; LEE, L.; TANG, J.; JIANG, L.; ZHOU, M.; LI, P. User-level sentiment analysis incorporating social networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2011. p. 1397–1405. Citado na página 33.

TANSCHEIT, R. Inteligência computacional: Aplicada a administração, economia e engenaria em matlab. In: *Sistemas Fuzzy*. [S.l.]: Thomson Pioneira, 2004. Citado 5 vezes nas páginas 19, 4, 15, 16 e 17.

TEMPLIN, M. *Certain Language Skills in Children: Their Development and Interrelationships*. [S.l.]: University of Minnesota Press, 1957. Citado na página 41.

THET, T. T.; NA, J.-C.; KHOO, C. S. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, v. 36, n. 6, p. 823–848, 2010. Citado na página 13.

TORRUELLA, J.; CAPSADA, R. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, v. 95, p. 447 – 454, 2013. Citado 2 vezes nas páginas 41 e 42.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Mining Knowledge Discovery*, v. 24, n. 3, p. 478–514, 2012. Citado na página 12.

TURNEY, P. D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. [S.l.: s.n.], 2002. p. 417–424. Citado 6 vezes nas páginas 21, 31, 32, 34, 38 e 39.

VASWANI, A.; ZHAO, Y.; FOSSUM, V.; CHIANG, D. Decoding with large-scale neural language models improves translation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2013. p. 1387–1392. Citado na página 1.

VINODHINI, G.; CHANDRASEKARAN, R. Sentiment analysis and opinion mining: a survey. *International Journal of Advanced Research in Computer Science and Software Engneering*, v. 2, n. 6, 2012. Citado na página 36.

VOHS, K. D.; BAUMEISTER, R. F.; SCHMEICHEL, B. J.; TWENGE, J. M.; NELSON, N. M.; TICE, D. M. Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology*, v. 94, p. 883–898, 2008. Citado na página 2.

WANG, D.; LIU, Y. Opinion summarization on spontaneous conversations. *Computer Speech and Language*, v. 34, n. 1, p. 61–82, 2015. Citado na página 35.

WANG, H.; LU, Y.; ZHAI, C. Latent aspect rating analysis on review text data: A rating regression approach. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2010. p. 783–792. Citado na página 36.

WARD, J.; PEPPARD, J. *The Strategic Management of Information Systems: Building a Digital Strategy*. [S.l.]: John Wiley & Sons, 2016. Citado na página 1.

WEBER, L.; KLEIN, P. *Aplicação da lógica fuzzy em software e hardware*. [S.l.]: Editora da ULBRA, 2003. Citado na página 46.

WIDROW, B.; HOFF, M. E. Adaptive switching circuits. In: *Neurocomputing: Foundations of Research*. [S.l.: s.n.], 1988. p. 123–134. Citado na página 21.

XIMENES, S. *Minidicionário da Língua Portuguesa*. 2nd. ed. [S.l.]: Ediouro, 2000. Citado na página 32.

YATES, F. Contingency tables involving small numbers and the x2 test. *Supplement to the Journal of the Royal Statistical Society*, v. 1, n. 2, p. 217–235, 1934. Citado na página 58.

ZADEH, L. A. Fuzzy sets. *Information and Control*, v. 8, p. 338–353, 1965. Citado na página 15.

ZADEH, L. A. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, v. 8, n. 3, p. 199–249, 1975. Citado na página 15.

ZADEH, L. A. Fuzzy logic and approximate reasoning. *Synthese*, v. 30, n. 3, p. 407–428, 1975. Citado na página 15.

ZHANG, L.; LIU, B. Aspect and entity extraction for opinion mining. In: *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities*. [S.l.]: Springer Berlin Heidelberg, 2014. p. 1–40. Citado na página 10.

ZHANG, P.; HE, Z. Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science*, v. 41, n. 4, p. 531–549, 2015. Citado na página 1.

Apêndices

APÊNDICE A – Base de Características

Característica	***Stem***	**Característica**	***Stem***	**Característica**	***Stem***
água quente	*no stem*	descanso	descans	pessoa	pesso
alcatifa	alcatif	elevador	elevad	piscina	piscin
apartamento	apart	equipe	equip	ponto	pont
ar-condicionado	ar-condicion	escuridão	escur	posição	posic
aroma	arom	estacionamento	estacion	preço	preç
atendimento	atend	estatuto	estatut	quarto	quart
bar	bar	estoque	estoqu	quarto de banho	*no stem*
banheiro	banh	faxina	faxin	recepção	recepç
barulho	barulh	fila	fil	regimento	reg
benefício	benefici	funcionário	funcionari	repouso	repous
cadeira	cade	gastronomia	gastronom	reserva	reserv
café da manhã	*no stem*	gerência	ger	restaurante	restaurante
cama	cam	hotel	hotel	saída	said
camareira	camar	instalação	instal	salão do café	*no stem*
carpete	carpet	internet	internet	salário	salári
carro	carr	lanche	lanch	serviço	serviç
casa de banho	*no stem*	lanterna	lantern	silêncio	silênc
cervejaria	cervej	leque	lequ	suíte	suíte
check	check	limpeza	limpeza	tamanho	tamanh
cheiro	cheir	localização	local	táxi	tax
chuveiro	chuv	lugar	lug	toalha	toalh
comodidade	comod	mesa	mes	trabalho	trabalh
conectividade	conect	móvel	móve	transporte	transport
conforto	confort	pagamento	pag	valor	val
conjunto	conjunt	paisagem	paisag	veículo	veícul
corredor	corredor	paróquia	paróqu	vista	vist
custo	cust	partida	partida	wifi	wifi

APÊNDICE B – Arquivo FCL

```
FUNCTION_BLOCK getImportance

     VAR_INPUT
          author: REAL;
          patt : REAL;
          richness : REAL;
     END_VAR

     VAR_OUTPUT
          importance: REAL;
     END_VAR

     FUZZIFY author
          TERM low := (0, 1) (3, 1) (4, 0) ;
          TERM medium := (3, 0) (4, 1) (7, 1) (9, 0) ;
          TERM high := (6, 0) (8, 1) (10, 1);
     END_FUZZIFY

     FUZZIFY patt
          TERM low := (0, 1) (3, 1) (4, 0) ;
          TERM medium := (3, 0) (4, 1) (7, 1) (9, 0) ;
          TERM high := (6, 0) (8, 1) (10, 1);
     END_FUZZIFY

     FUZZIFY richness
          TERM bad := (0, 1) (0.5, 1) (1, 0) ;
          TERM medium := (2, 0) (3, 1) (5, 1) (6, 0) ;
          TERM good := (5, 0) (6, 1) (8, 1) (9, 0) ;
          TERM excellent := (2.5, 0) (3, 1) (4, 1) ;
     END_FUZZIFY

     DEFUZZIFY importance
          TERM isf := (0, 1) (2, 1) (3, 0);
          TERM sf := (2, 0) (3.5, 1) (5, 0);
          TERM good := (5, 0) (6.5, 1) (8, 0);
```

```
    TERM exc := (8, 0) (9, 1) (10, 1);

    METHOD : COG ;
    DEFAULT := 0 ;
END_DEFUZZIFY

RULEBLOCK first

    AND : MIN;
    ACCU : MAX;

    RULE 1 : IF (author IS low) AND (patt IS low)
    AND (richness IS bad) THEN (importance IS isf) ;

    RULE 2 : IF (author IS low) AND (patt IS low)
    AND (richness IS medium)THEN (importance IS isf) ;

    RULE 3 : IF (author IS low) AND (patt IS low)
    AND (richness IS good) THEN (importance IS isf) ;

    RULE 4 : IF (author IS low) AND (patt IS low)
    AND (richness IS excellent) THEN (importance IS sf) ;

    RULE 5 : IF (author IS low) AND (patt IS medium)
    AND (richness IS bad) THEN (importance IS isf) ;

    RULE 6 : IF (author IS low) AND (patt IS medium)
    AND (richness IS medium)THEN (importance IS sf) ;

    RULE 7 : IF (author IS low) AND (patt IS medium)
    AND (richness IS good) THEN (importance IS sf) ;

    RULE 8 : IF (author IS low) AND (patt IS medium)
    AND (richness IS excellent) THEN (importance IS good) ;

    RULE 9 : IF (author IS low) AND (patt IS high)
    AND (richness IS bad) THEN (importance IS sf) ;

    RULE 10 : IF (author IS low) AND (patt IS high)
```

```
AND (richness IS medium) THEN (importance IS sf) ;

RULE 11 : IF (author IS low) AND (patt IS high)
AND (richness IS good) THEN (importance IS good) ;

RULE 12 : IF (author IS low) AND (patt IS high)
AND (richness IS excellent) THEN (importance IS good) ;

RULE 13 : IF (author IS medium) AND (patt IS low)
AND (richness IS bad) THEN (importance IS isf) ;

RULE 14 : IF (author IS medium) AND (patt IS low)
AND (richness IS medium) THEN (importance IS isf) ;

RULE 15 : IF (author IS medium) AND (patt IS low)
AND (richness IS good) THEN (importance IS isf) ;

RULE 16 : IF (author IS medium) AND (patt IS low)
AND (richness IS excellent) THEN (importance IS sf) ;

RULE 17 : IF (author IS medium) AND (patt IS medium)
AND (richness IS bad) THEN (importance IS sf) ;

RULE 18 : IF (author IS medium) AND (patt IS medium)
AND (richness IS medium) THEN (importance IS sf) ;

RULE 19 : IF (author IS medium) AND (patt IS medium)
AND (richness IS good) THEN (importance IS good) ;

RULE 20 : IF (author IS medium) AND (patt IS medium)
AND (richness IS excellent) THEN (importance IS good) ;

RULE 21 : IF (author IS medium) AND (patt IS high)
AND (richness IS bad) THEN (importance IS sf) ;

RULE 22 : IF (author IS medium) AND (patt IS high)
AND (richness IS medium) THEN (importance IS sf);

RULE 23 : IF (author IS medium) AND (patt IS high)
```

```
AND (richness IS good) THEN (importance IS good);

RULE 24 : IF (author IS medium) AND (patt IS high)
AND (richness IS excellent) THEN (importance IS exc);

RULE 25 : IF (author IS high) AND (patt IS low)
AND (richness IS bad) THEN (importance IS sf) ;

RULE 26 : IF (author IS high) AND (patt IS low)
AND (richness IS medium) THEN (importance IS sf) ;

RULE 27 : IF (author IS high) AND (patt IS low)
AND (richness IS good) THEN (importance IS sf) ;

RULE 28 : IF (author IS high) AND (patt IS low)
AND (richness IS excellent) THEN (importance IS good) ;

RULE 29 : IF (author IS high) AND (patt IS medium)
AND (richness IS bad) THEN (importance IS sf) ;

RULE 30 : IF (author IS high) AND (patt IS medium)
AND (richness IS medium) THEN (importance IS good) ;

RULE 31 : IF (author IS high) AND (patt IS medium)
AND (richness IS good) THEN (importance IS good) ;

RULE 32 : IF (author IS high) AND (patt IS medium)
AND (richness IS excellent) THEN (importance IS exc) ;

RULE 33 : IF (author IS high) AND (patt IS high)
AND (richness IS bad) THEN (importance IS good) ;

RULE 34 : IF (author IS high) AND (patt IS high)
AND (richness IS medium) THEN (importance IS exc) ;

RULE 35 : IF (author IS high) AND (patt IS high)
AND (richness IS good) THEN (importance IS exc) ;

RULE 36 : IF (author IS high) AND (patt IS high)
```

```
        AND (richness IS excellent) THEN (importance IS exc) ;

    END_RULEBLOCK

END_FUNCTION_BLOCK getImportance
```

Anexos

ANEXO A – Etiquetas Mac-Morpho

Classe Gramatical	Etiqueta	Exemplos
Adjetivo	**ADJ**	bom - ruim - ótimo - péssimo
Advérbio	**ADV**	muito - pouco - normalmente
Advérbio Conectivo Subordinativo	**ADV-KS**	Sei onde mora
Advérbio Relativo Subordinativo	**ADV-KS-REL**	onde - quando - como
Artigo	**ART**	o - a - os - as
Conjunção Coordenativa	**KC**	e - nem - mas - ou - pois
Conjunção Subordinativa	**KS**	que - porque - assim
Interjeição	**IN**	ufa! - viva! - ai! - oi!
Numeral	**NUM**	três - quatro - 3 - 4
Palavra Denotativa	**PDEN**	até - apenas - eis - cá
Particípio	**PCP**	dormido - espalhado - tido
Pronome Adjetivo	**PROADJ**	meu - nosso - este - algum
Pronome Conectivo Subordinativo	**PRO-KS**	Sei quem chegou
Pronome Conectivo Subord. Relativo	**PRO-KS-REL**	o qual - cujo
Pronome Pessoal	**PROPESS**	eu - me - Vossa Alteza
Pronome Substantivo	**PROSUB**	isto - isso - aquilo - alguém
Símbolo de Moeda Corrente	**CUR**	R$ - US$
Substantivo	**N**	hotel - quarto - atendimento
Substantivo Próprio	**NPROP**	Maria - Vinícius - Globo
Verbo	**V**	é - foi - gostar - ir
Verbo Auxiliar	**VAUX**	ter - haver