

A Comparative Study among Approaches based in Fuzzy Systems and Artificial Neural Networks to Estimate Importance of Comments about Products and Services

Roney Lira de Sales Santos
Computer Science Department
Federal University of Piauí
Teresina, Piauí, Brazil
E-mail: roneylira@hotmail.com

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 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. 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 information and others measures. Sousa [1] proposed an approach, called TOP(X), to estimate the degree of importance of reviews 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 work 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.

Index Terms—Artificial Neural Networks, Fuzzy Systems, Natural Language Processing, Opinion Mining.

I. INTRODUÇÃO

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 [2]. 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 [3].

Segundo [4], 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. [4] 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 [5]–[7].

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 e-mails, pesquisas, *posts* em redes sociais e fóruns, entre outros¹. 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 anos². 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 [4].

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*) [8]. A mineração de dados, segundo [9] 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.

Um dos domínios existentes no *e-commerce* utilizado pela mineração de dados é o de hotéis, onde de acordo com [10], 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 leitores³. O *Booking.com* é um *website* de reserva de hotéis disponível em 40 idiomas e ativo em 226 países e territórios⁴. No Brasil, o *Booking.com* tem 32.542 hotéis cadastrados⁵. 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.

Como existe um grande número de comentários publicados pelos hóspedes, 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 [11]: 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.

A. Objetivos

Com o objetivo de identificar os comentários mais relevantes, [1] 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. [12], [13].

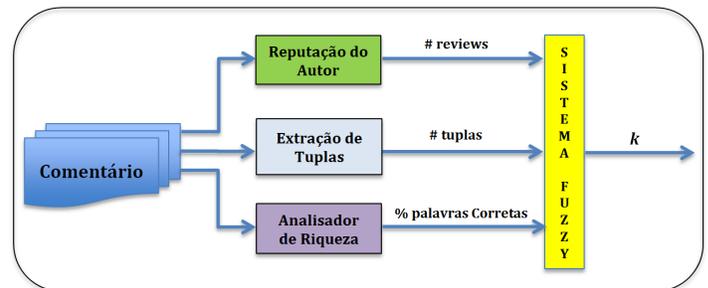


Figura 1. Estrutura geral do Sistema *Fuzzy* por [1].

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, neste trabalho, 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 *Cópus* 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;

¹Disponível em <http://www.clarabridge.com/nlp-natural-language-processing/>

²http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf

³Disponível em <http://www.melhoresdestinos.com.br/melhor-site-reservar-hotels.html>

⁴Disponível em <http://www.booking.com/content/about.pt-br.html>

⁵Disponível em <https://www.booking.com/country.pt-br.html>. Dados de dezembro/2016

- 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 Aqui⁶.

B. 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;
- 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 *Corpus*⁷ 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* [14]: 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 [1] com uma abordagem utilizando RNA com as mesmas variáveis de entrada no mesmo domínio de aplicação. Nesse trabalho, a abordagem com Sistema *Fuzzy* de [1] 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 neste trabalho.
- *Extração de Métricas e Análise de Sentimentos em Comentários Web no Domínio de Hotéis* [15]: 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 neste trabalho. 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* [16]: Este trabalho apresenta uma parte da segunda versão do trabalho proposto neste trabalho, 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 [17] e Maas, se juntando à correção 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.

C. Organização

O restante do artigo está organizado como: A Seção II são apresentadas as áreas de pesquisa deste trabalho. Na Seção III é apresentada uma revisão da literatura englobando os principais trabalhos da área. A Seção IV descreve 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. Na Seção V são descritos os experimentos realizados em cada uma das abordagens propostas neste trabalho, detalhes de cada execução e a comparação entre as abordagens. Por fim, a Seção VI apresenta as conclusões, limitações e os trabalhos futuros propostos para continuação da pesquisa.

II. REFERENCIAL TEÓRICO

A. Mineração de Opinião

De forma científica, [4] define uma opinião como sendo uma quintupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ 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 [4]. Todos os componentes da quintupla 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

⁶<http://www.reclameaqui.com.br/>

⁷Foi utilizada esta nomenclatura para indicar a coleção de documentos (comentários), que indicará os termos *corpus* e *corpora* usados por outros pesquisadores.

para um tipo de opinião chamada opinião regular, diferente da opinião comparativa que necessita de uma definição diferente [18].

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 [19], pode ser no nível de: Documento, Sentença e Entidade ou Aspeto. [19] ainda define os tipos de opiniões como sendo: Regulares (Diretas ou Indiretas [20]) ou Comparativas e Explícitas ou Implícitas.

B. Sistemas Fuzzy

A Teoria de Conjuntos *Fuzzy* [21] e os conceitos de Lógica *Fuzzy* [22], [23] 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 [12].

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* [24].

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 :

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

O conjunto *Fuzzy* proposto por [21] 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 [12].

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 [25]. O modelo de um sistema de inferência *Fuzzy* é mostrado na Figura 2.

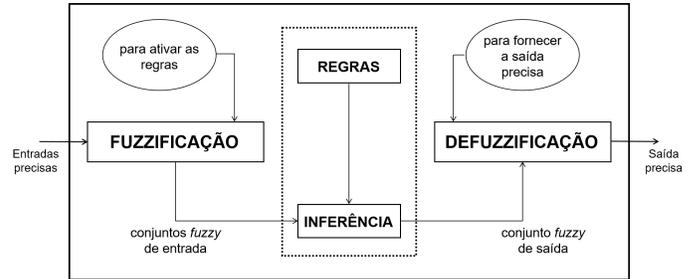


Figura 2. Modelo de Sistema de Inferência *Fuzzy* [12]

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*. 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 [12].

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 [26] 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.

C. Redes Neurais Artificiais

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 [27]. 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 [28].

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 [29]. 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 [30] 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 3 mostra o modelo de neurônio artificial. O funcionamento de um neurônio artificial começa pela apresentação dos sinais de entrada representados pelo conjunto $\{x_1, x_2, \dots, x_n\}$ que são ponderados por meio do conjunto de pesos sinápticos $\{w_1, w_2, \dots, w_n\}$. 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 $g(\cdot)$ 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.

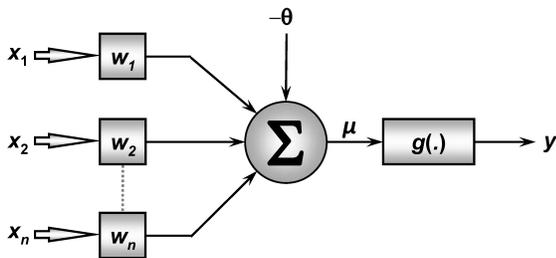


Figura 3. Neurônio artificial [28]

Em [31] 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.

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: *feedforward* de camada simples, *feedforward* de múltiplas camadas, recorrentes ou reticuladas. Entre os principais tipos de redes tendo arquitetura *feedforward* de camada simples estão o *Perceptron* [32] e o *Adaline* [33], cujos algoritmos de aprendizado utilizados em seus treinamentos são os baseados na regra de Hebb [34] e na regra Delta [33]. 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) [35]. Já as principais redes com arquitetura recorrente são as redes de Hopfield [36] e a rede *Perceptron* com realimentação. Por fim, a rede de Kohonen [37] é a principal representante das redes com arquitetura reticulada.

III. TRABALHOS RELACIONADOS

A mineração de opinião consiste em três etapas [38]: 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.

A. 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 [39] define cinco padrões linguísticos utilizados na extração de frases contendo adjetivos e verbos. Essas classes gramaticais demonstram ser bons indicadores de subjetividade e sentenças avaliativas

para definir a classificação do sentimento [40]. Os padrões linguísticos, para o inglês, são mostrados na Tabela I. Uma adaptação desses padrões foi feita para serem utilizados neste trabalho, a qual será explicada em detalhes na seção IV.

Tabela I
PADRÕES DE TURNEY [39]

Padrão	1ª 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 [41] 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.

Já Silva, Lima e Barros [42] 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 [39].

B. 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% [43], [44].

As principais técnicas de classificação são: i) uso de léxico (dicionário) com palavras ou expressões de sentimentos, por exemplo, *Sentilex-PT* [45] e *OpLexicon* [46]; e ii) aprendizagem de máquina, por exemplo, Máquinas de Vetores de Suporte (do inglês *Support Vector Machine*, SVM) [47], [48], Sistemas *Fuzzy* [1], [49], [50] e Redes Neurais Artificiais [51], [52].

Em relação à orientação semântica por meio de léxico de sentimentos, Kamps et al. [53] utilizaram as relações semânticas presentes na *WordNet*. Um grafo com os adjetivos contidos na interseção 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.

Santos et al. [14] realizaram um estudo experimental que envolvia a abordagem de Sousa [1] utilizando Sistemas *Fuzzy* com uma abordagem similar, porém utilizando uma Rede Neural Artificial. Os mesmos experimentos definidos em Sousa [1] foram realizados a fim de obter uma comparação justa entre as abordagens. Uma RNA MLP com o algoritmo de aprendizagem *backpropagation* [54] 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 neste trabalho.

C. 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.

A nível de entidade ou aspecto, Kurian e Asokan [55] 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* [56]. 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, Kokelar et al. [57] 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 [58] usaram o método de pontuação de frequência e relevância de termos para representar as frases mais informativas na sumarização. Já Wang, Lu e Zhai [59] 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. [60] e uma comparação de métodos de sumarização de opiniões é tratado por Condori e Pardo [61]. Outras pesquisas (*surveys*) que contém mineração de opinião e métricas para sumarização da informação são encontradas em [62] e [63].

IV. 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 [1], 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 neste trabalho.

A. Abordagem TOP(X) Original

[1] 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 correteza 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 neste trabalho de **abordagem base**.

Para uma comparação justa com as abordagens propostas neste trabalho, 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.

B. Adaptações

As variáveis utilizadas por [1] 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.

1) *Adaptações na variável quantidade de tuplas*: [1] 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 [39] 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 características⁸, 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* [64] para coletar os sinônimos das características previamente encontradas. Nele foram encontradas mais 28 características. Uma filtragem manual foi realizada a fim de eliminar as características que continham o mesmo *stemming*⁹, finalizando a base de características com 81 palavras. Esta base de características está livremente disponível para utilização¹⁰.

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 neste trabalho. Ainda com base nos padrões de [39], novos padrões linguísticos foram propostos para a extração das tuplas da abordagem. A Tabela II mostra os padrões propostos.

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.

O padrão 4 verifica se junto das características se encontram verbos e advérbios. A identificação de verbos é importante

pois segundo [65], 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 neste trabalho é mostrada na Tabela III. 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.

Verifica-se na Tabela III 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 IV 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.

2) *Adaptações na variável riqueza de vocabulário*: Sousa [1] 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*) [66], o índice de Maas [67] e o índice MTL (do inglês: *Measure of Textual Lexical Diversity*) [68].

[69] 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 (1)

$$TTR = \frac{t}{n} \quad (1)$$

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 (2)

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

⁸Neste trabalho considera-se como característica as entidades, aspectos e seus atributos.

⁹Termo usado na linguística para descrever o processo de redução (flexionado ou derivado) da palavra.

¹⁰Disponível em <https://goo.gl/SAe9Rr>

Tabela II
PADRÕES LINGÜÍSTICOS 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>)

Tabela III
OCORRÊNCIA DOS PADRÕES LINGÜÍSTICOS PROPOSTOS

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

Tabela IV
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%

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 MTLT. O cálculo do índice MTLT 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 [70], 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 MTLT é feito da seguinte maneira: of (1,00) the (1,00) people (1,00) by (1,00) the (0,80) people (0,667) || $s = s + 1$ || 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 [70]. 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) [71]

$$SP = \frac{1 - TTR_{lastword}}{1 - 0,72} \quad (3)$$

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 MTLT é calculado pela Equação (4)

$$MTLT = \frac{n}{s + SP} \quad (4)$$

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 MTLT é feita com o texto sendo processado de forma inversa. A média dos dois valores é o valor final do índice MTLT [70].

Em relação ao tamanho do texto, [69] 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 MTLT.

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 MTLT 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 [1], as quatro variáveis propostas neste trabalho para definir a riqueza do vocabulário devem ser convertidas em apenas um valor de entrada. A justificativa para tal vem de [1] 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 4.

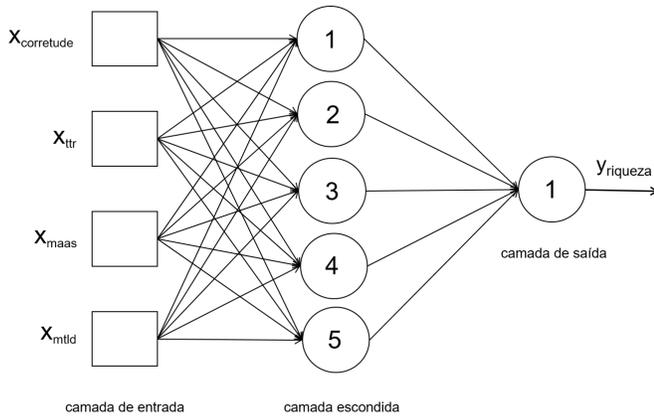


Figura 4. Topologia: RNA riqueza do vocabulário

A topologia da RNA foi definida por meio do *software* SPSS [72]. 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*) [73].

Para a realização do treinamento e teste foram considerados 370 comentários que foram manualmente classificados como Ruim (RM), Médio (MD), Bom (BM) e Excelente (EXC) com referência aos valores de corretude, TTR, Maas e MTLID. 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 V. 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%.

Tabela V
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

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

C. Abordagem com Sistema *Fuzzy*

Levando em conta as três variáveis de entrada propostas por [1] e as adaptações propostas em duas delas, quantidade

de tuplas e riqueza do vocabulário, discutidas nas seções IV-B1 e IV-B2, o esquema geral da abordagem com Sistema *Fuzzy* é apresentado na Figura 5.

A abordagem utiliza o modelo de inferência de Mandami [26]. 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 [1].

Como explicado na Seção II-B, 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 VI.

Tabela VI
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 [1]. Foi necessário adaptar a variável riqueza do vocabulário pois, como explicado na Seção IV-B2, 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 6: reputação do autor (Figura 6a), quantidade de tuplas (Figura 6b), riqueza do vocabulário (Figura 6c) e da importância do comentário (Figura 6d). É importante mencionar que estas funções de pertinência não são similares às funções de pertinência propostas por [1], pois houveram alterações no comportamento da variável riqueza do vocabulário, fazendo com que as funções de pertinência fossem revisadas.

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

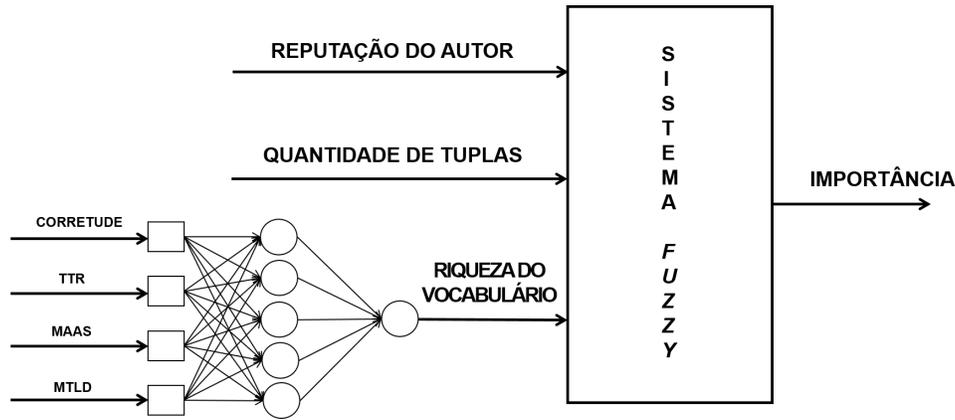


Figura 5. Abordagem com Sistema *Fuzzy*

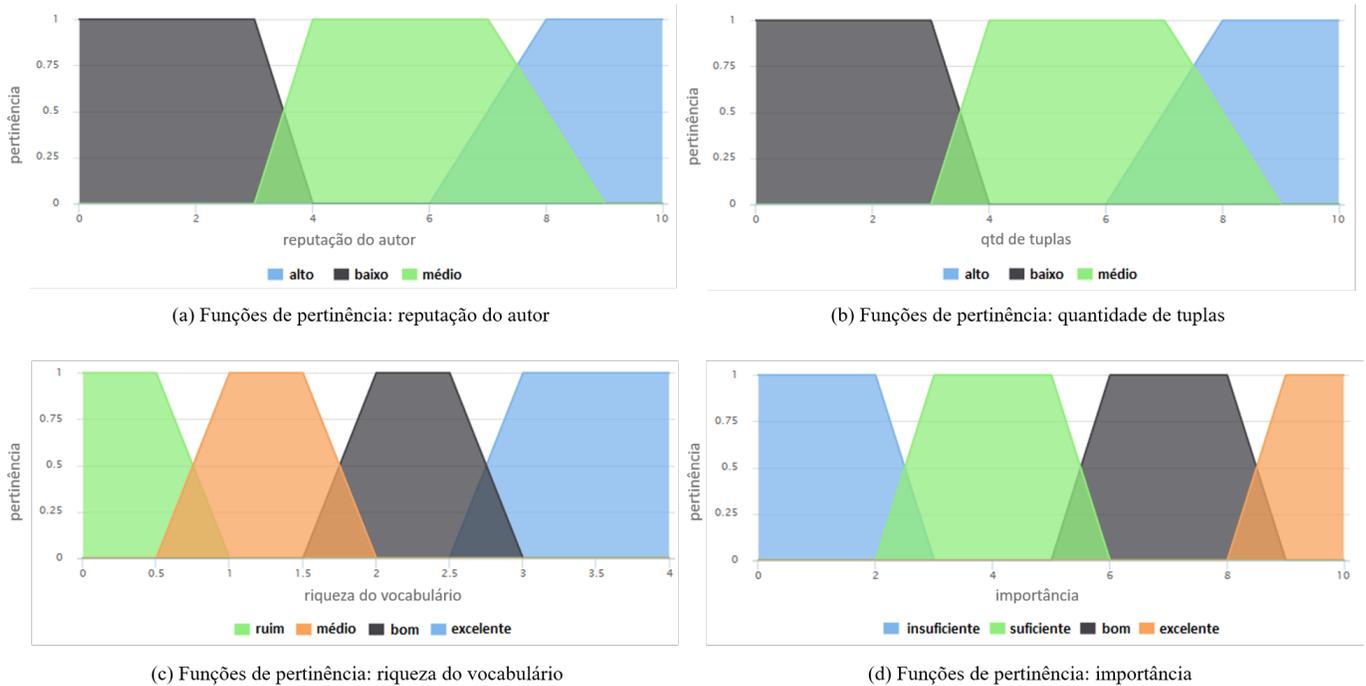


Figura 6. Funções de pertinência

mostrada na Tabela VII. 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.

Como exemplo de leitura das regras da Tabela VII, 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 cen-

tro de área (centro de gravidade¹¹ 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 [74]. Sua equação matemática é definida pela Equação (5)

$$Y = \frac{\sum_{i=1}^{qc} \mu_i \cdot Y_i}{\sum_{i=1}^{qc} \mu_i} \quad (5)$$

¹¹Na maioria dos casos, o centro da área está na mesma posição do centro de gravidade, então estes nomes frequentemente denotam o mesmo método [74].

Tabela VII
BASE DE REGRAS DO SISTEMA *Fuzzy*

Autor	Quantidade de Tuplas/Riqueza do Vocabulário											
	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

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 [75]. A execução da abordagem foi realizada na ferramenta Athena.

D. Abordagem com Rede Neural Artificial

No decorrer da seção II-C 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 única¹² 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 [28], [31], [78].

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 [28]. Por outro lado, o treinamento das redes RBF geralmente é mais rápido que o treinamento das redes MLP [79].

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 neste trabalho 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*¹³.

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 [80], 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 7 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*).

¹²É comum associar a rede RBF a uma rede com uma camada escondida, embora redes RBF com mais de uma camada escondida tenham sido propostas [76], [77]

¹³Disponível em https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_mlp_architecture_expert.htm

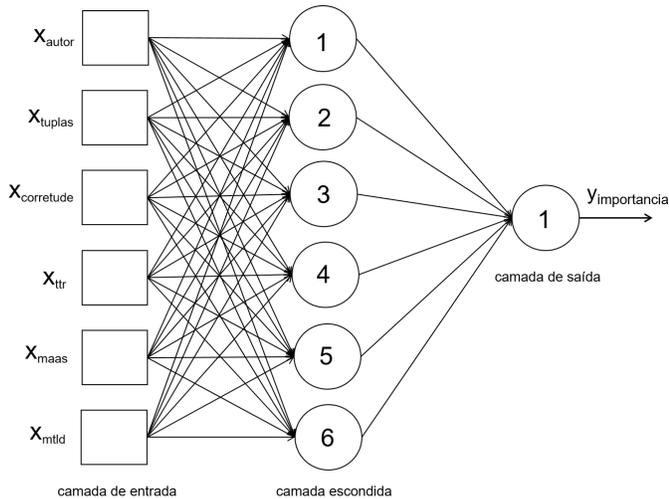


Figura 7. Topologia: RNA Multi-Layer Perceptron

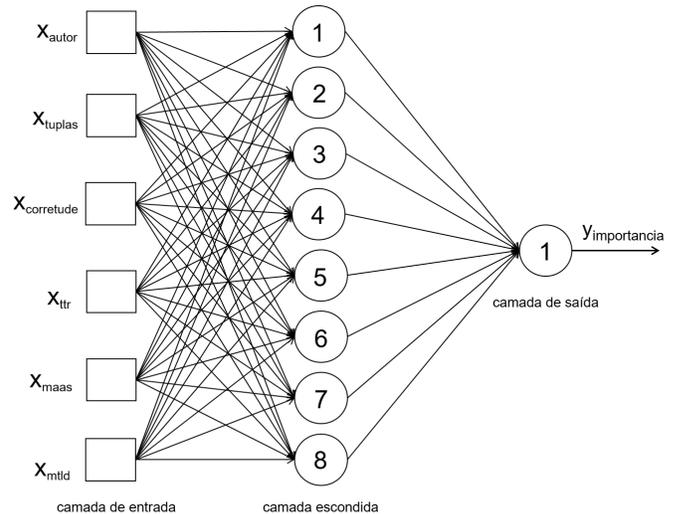


Figura 8. Topologia: Redes de Base Radial

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*¹⁴ 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 8, 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 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) [81], [82], a qual utiliza a função *softmax* [83]. A função do erro utilizada nesta topologia foi a soma dos erros quadrados [84].

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 MTLT. Diferente do que proposto no Sistema *Fuzzy* explicado na Seção IV-C, 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.

V. EXPERIMENTOS E DISCUSSÕES

Nesta seção são discutidos os resultados referentes às abordagens propostas neste trabalho. 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.

A. Resultados

Com o objetivo de comparar as abordagens, foi realizado um experimento com o subcorpus 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 [85].

A Tabela VIII 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 IX.

¹⁴Disponível em https://www.ibm.com/support/knowledgecenter/SSLVMB_20.0.0/com.ibm.spss.statistics.help/alg_rbf_training_autoselect_numhiddenunits.htm

Tabela VIII
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

Tabela IX
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%

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

$$P = \frac{TP}{(TP + FP)} \quad (6)$$

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

$$R = \frac{TP}{(TP + FN)} \quad (7)$$

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 [86]. O cálculo da medida-F é realizado segundo a equação

$$F = \frac{2 * P * R}{P + R} \quad (8)$$

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 [87], mas neste trabalho 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 X apresenta a matriz de confusão para a abordagem com Sistema *Fuzzy* e adaptações e a Tabela XI mostra as medidas de precisão, cobertura, medida-F e taxa de erro por classe. A precisão total da abordagem foi de 60,54%.

Tabela X
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 XI
MEDIDAS DE AVALIAÇÃO: ABORDAGEM COM SISTEMA *Fuzzy* E ADAPTAÇÕES

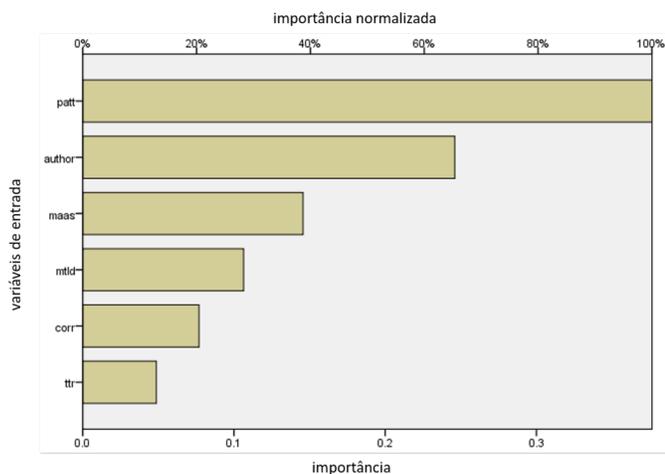
Classe	P	R	F	Taxa de Erro
EXC	92,8%	65%	76,4%	35%
BM	72,1%	44,8%	55,3%	55,2%
SF	49,5%	45,5%	47,4%	54,5%
ISF	60,9%	86%	71,3%	14%

A matriz de confusão referente à abordagem com RNA tipo MLP é mostrada na Tabela XII e na Tabela XIII 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%.

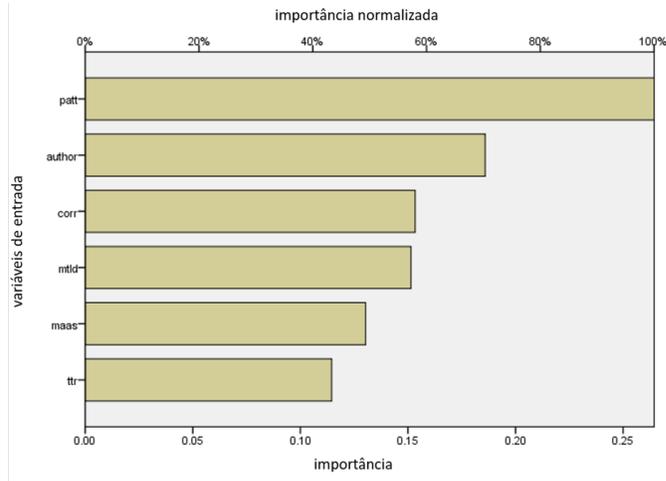
Na Tabela XIV é apresentada a matriz de confusão referente à abordagem com RNA tipo RBF, enquanto que as medidas de precisão, cobertura, medida-F e taxa de erro por classe são mostradas na Tabela XV. Sua precisão total foi de 64,32%.

B. 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



(a) Importância das variáveis: RNA MLP



(a) Importância das variáveis: RNA RBF

Figura 9. Funções de pertinência

Tabela XII
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 XIII
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%

Tabela XIV
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 XV
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%

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 9a e 9b, respectivamente.

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 [88]. Observa-se nas figuras 9a e 9b 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 IV-B2), 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 [89], [90]; ii) Eliminação Recursiva de Características (do inglês: *Recursive Feature Elimination*, RFE) [91]; 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 XVI.

Tabela XVI
IMPORTÂNCIA DAS VARIÁVEIS: POR TÉCNICAS DE SELEÇÃO DE CARACTERÍSTICAS

Seleção Univariante		RFE		ADA	
1. Tuplas	0,403	1. Tuplas	TRUE	1. Tuplas	0,304
2. Autor	0,372	2. Autor	TRUE	2. Autor	0,164
3. MTLT	0,208	3. Maas	FALSE	3. Maas	0,143
4. Corretude	0,010	4. Corretude	FALSE	4. MTLT	0,135
5. TTR	0,004	5. MTLT	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 MTLT. 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).

C. 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 neste trabalho. Como as abordagens que utilizam Sistema

Fuzzy e Redes Neurais Artificiais são paramétricas [92], fez-se uma comparação cuidadosamente planejada. As Seções IV-C e IV-D especificam os parâmetros de cada abordagem.

Para facilitar a comparação, a Tabela XVII mostra as precisões totais (acurácia) das quatro abordagens estudadas neste trabalho. 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 XVII
PRECISÃO TOTAL DE CADA ABORDAGEM

Abordagem	Acurácia
Original	52,7%
Sistema <i>Fuzzy</i> 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 aceita pela comunidade, pois ela é considerada uma medida fraca [93], além de que, um modelo cujo objetivo é maximizar a acurácia pode aparentemente ter uma boa avaliação, pois pode considerar as informações irrelevantes [94]. 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 XVIII.

Tabela XVIII
MEDIDA-F POR CLASSE DE CADA ABORDAGEM

Classe	Original	Sistema <i>Fuzzy</i> 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 [95], 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

$$\varphi = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (9)$$

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 neste trabalho, sendo geralmente considerada como sendo uma das melhores métricas para se medir a performance de um sistema [85], 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 XIX.

Tabela XIX
COEFICIENTE DE CORRELAÇÃO DE MATTHEWS POR CLASSE DE CADA ABORDAGEM

Classe	Original	Sistema <i>Fuzzy</i> 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

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 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 XX 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).

Verifica-se que em todas 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:

Tabela XX
MEDIDA-F POR POLARIDADE DE CADA CLASSE DE CADA ABORDAGEM

Classe	Original	Sistema <i>Fuzzy</i> 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%

- 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 neste trabalho, 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 *Cópus* de comentários sobre hotéis é utilizando Sistema *Fuzzy*, enquanto a abordagem que melhor identifica os comentários mais importantes de um *Cópus* de comentários sobre hotéis anotado com sua orientação semântica (positivo ou negativo) é utilizando Redes Neurais Artificiais MLP.

VI. 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 [1]. 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 correteza do comentário, a adaptação proposta passou pela implementação de uma RNA na qual as entradas seriam os valores da correteza 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 [1], 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 das RNAs 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 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 *Corpus* 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 [96]; 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*¹⁵ 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.

A. 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 neste trabalho 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;
- Propor uma nova abordagem utilizando Sistemas Neuro-Fuzzy, a qual se integram os dois modelos computacionais explorados neste trabalho, 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 [98];
- 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 [99].

REFERÊNCIAS

- [1] R. F. de Sousa, "Abordagem top(x) para inferir comentários mais importantes sobre produtos e serviços," Master's thesis, Universidade Federal do Piauí, 2015.
- [2] M. Lousada and M. L. Pomim Valentim, "Modelos de tomada de decisão e sua relação com a informação orgânica," *Perspectivas em Ciência da Informação*, vol. 16, pp. 147–164, 2011.
- [3] J. Ward and J. Peppard, *The Strategic Management of Information Systems: Building a Digital Strategy*. John Wiley & Sons, 2016.

¹⁵O *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". [97]

- [4] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.
- [5] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes, "Social network analysis and mining for business applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.
- [6] H. Gil de Zúñiga, "Social media use for news and individuals' social capital, civic engagement and political participation," *Journal of Computer-Mediated Communication*, vol. 17, no. 3, pp. 319–336, 2012.
- [7] D. Milne and I. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, vol. 194, pp. 222–239, 2013.
- [8] G. Piatetsky-Shapiro, "Knowledge discovery in real databases: A report on the IJCAI-89 workshop," *AI Magazine*, vol. 11, no. 5, pp. 68–70, 1991.
- [9] M. J. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., 1997.
- [10] W. Kasper and M. Vela, "Sentiment analysis for hotel reviews," in *Proceedings of the Computational Linguistics-Applications Conference*, 2011, pp. 45–52.
- [11] M. Li, L. Huang, C.-h. Tan, and K. K. Wei, "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce*, vol. 17, no. 4, pp. 101–136, 2013.
- [12] R. Tanscheit, "Inteligência computacional: Aplicada a administração, economia e engenharia em matlab," in *Sistemas Fuzzy*. Thomson Pioneira, 2004.
- [13] S. Haykin, *Neural Networks and Learning Machines*. Prentice Hall, 2009.
- [14] R. L. de S. Santos, R. F. de Sousa, R. A. L. Rabelo, and R. S. Moura, "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)*, 2016, pp. 647–653.
- [15] R. L. de S. Santos and R. S. Moura, "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*, 2016, pp. 127–138.
- [16] R. L. de S. Santos, J. ao Paulo Albuquerque Vieira, J. L. N. Barbosa, C. A. Sá, E. G. Moura, R. S. Moura, and R. F. de Sousa, "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*, 2016, pp. 153–163.
- [17] J. B. Carroll, *Language and thought*. Prentice-Hall Englewood Cliffs, 1964.
- [18] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 244–251.
- [19] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan and Claypool, 2012.
- [20] —, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Secaucus, USA: Springer-Verlag New York, Inc., 2006.
- [21] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.
- [22] —, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.
- [23] —, "Fuzzy logic and approximate reasoning," *Synthese*, vol. 30, no. 3, pp. 407–428, 1975.
- [24] W. Pedrycz and F. Gomide, *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Wiley-IEEE Press, 2007.
- [25] Y. Bai and D. Wang, "Fundamentals of fuzzy logic control—fuzzy sets, fuzzy rules and defuzzifications," in *Advanced Fuzzy Logic Technologies in Industrial Applications*. Springer, 2006, pp. 17–36.
- [26] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975.
- [27] R. Rojas, *Neural Networks: A Systematic Introduction*. Springer-Verlag New York, Inc., 1996.
- [28] I. N. da Silva, D. H. Spatti, and R. A. Flauzino, *Redes Neurais Artificiais: para engenharia e ciencias aplicadas*. Artliber, 2010.
- [29] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, no. 1, pp. 500–544, 1952.
- [30] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [31] D. Kriesel, *A Brief Introduction to Neural Networks*, 2007. [Online]. Available: <http://www.dkriesel.com>
- [32] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, pp. 65–386, 1958.
- [33] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in *Neurocomputing: Foundations of Research*, 1988, pp. 123–134.
- [34] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Wiley, 1949.
- [35] D. S. Broomhead and D. Lowe, "Radial basis functions, multi-variable functional interpolation and adaptive networks," DTIC Document, Tech. Rep., 1988.
- [36] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
- [37] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.
- [38] M. Tsytasarau and T. Palpanas, "Survey on mining subjective data on the web," *Data Mining Knowledge Discovery*, vol. 24, no. 3, pp. 478–514, 2012.
- [39] P. D. Turney, "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*, 2002, pp. 417–424.
- [40] V. Hatzivassiloglou and J. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of 18th International Conference on Computational Linguistics*, 2000, pp. 299–305.
- [41] J. Liu, G. Wu, and J. Yao, "Opinion searching in multi-product reviews," in *6th IEEE International Conference on Computer and Information Technology*, 2006, pp. 25–26.
- [42] N. R. Silva, D. Lima, and F. Barros, "Sapair: Um processo de análise de sentimento no nível de característica," in *4th International Workshop on Web and Text Intelligence*, 2012.
- [43] R. F. Bruce and J. M. Wiebe, "Recognizing subjectivity: A case study in manual tagging," *Natural Language Engineering*, vol. 5, no. 2, pp. 187–205, 1999.
- [44] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, 2002, pp. 79–86.
- [45] M. J. Silva, P. Carvalho, and L. Sarmento, "Building a sentiment lexicon for social judgement mining," in *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*, 2012, pp. 218–228.
- [46] M. Souza, R. Vieira, D. Buseti, R. Chishman, and I. M. Alves, "Construction of a portuguese opinion lexicon from multiple resources," in *In 8th Brazilian Symposium in Information and Human Language Technology - STIL*, 2011.
- [47] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2004, pp. 412–418.
- [48] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1397–1405.
- [49] M. Dragoni, A. G. B. Tettamanzi, and C. Da Costa Pereira, "A fuzzy system for concept-level sentiment analysis," in *Semantic Web Evaluation Challenge*, vol. 475, 2014, pp. 21–27.
- [50] K. Indhuja and R. P. C. Reghu, "Fuzzy logic based sentiment analysis of product review documents," in *First International Conference on Computational Systems and Communications (ICCS)*, 2014, pp. 18–22.
- [51] A. Sharma and S. Dey, "An artificial neural network based approach for sentiment analysis of opinionated text," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, 2012, pp. 37–42.
- [52] —, "A document-level sentiment analysis approach using artificial neural network and sentiment lexicons," *SIGAPP Applied Computing Review*, vol. 12, no. 4, pp. 67–75, 2012.
- [53] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke, "Using wordnet to measure semantic orientations of adjectives," in *In Proceedings of 4th*

- international conference on language resources and evaluation, 2004, pp. 1115–1118.
- [54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1986, pp. 318–362.
- [55] N. Kurian and S. Asokan, “Summarizing user opinions: A method for labeled-data scarce product domains,” *Procedia Computer Science*, vol. 46, pp. 93–100, 2015.
- [56] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” in *Proceedings of the 5th Conference on Language Resources and Evaluation*, 2006, pp. 417–422.
- [57] N. Kolekar, G. Rao, S. Dey, M. Mane, V. Jadhav, and S. Patil, “Sentiment analysis and classification using lexicon-based approach and addressing polarity shift problem,” *Journal of Theoretical and Applied Information Technology*, vol. 90, no. 1, pp. 118–125, 2016.
- [58] V. Raut and D. Londhe, “Opinion mining and summarization of hotel reviews,” in *Proceedings of the 6th International Conference on Computational Intelligence and Communication Networks*, 2014, pp. 556–559.
- [59] H. Wang, Y. Lu, and C. Zhai, “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*, 2010, pp. 783–792.
- [60] M. McNeill, R. Raeside, M. Graham, and I. Roseboom, “Comparing summarisation techniques for informal online reviews,” in *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 2015, pp. 322–329.
- [61] R. E. L. Condori and T. A. S. Pardo, “Opinion summarization methods: Comparing and extending extractive and abstractive approaches,” *Expert Systems with Applications*, vol. 78, pp. 124–134, 2017.
- [62] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: a survey,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 6, 2012.
- [63] V. Raut and D. Londhe, “Survey on opinion mining and summarization of user reviews on web,” *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 1026–1030, 2014.
- [64] H. Liu and P. Singh, “Conceptnet: A practical commonsense reasoning tool-kit,” *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, 2004.
- [65] J. Klavans and M.-Y. Kan, “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*, 1998, pp. 680–686.
- [66] M. Templin, *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, 1957.
- [67] H. D. Mass, “Zusammenhang zwischen wortschatzumfang und lange eines textes,” in *Zeitschrift für Literaturwissenschaft und Linguistik*, 1972, pp. 73–79.
- [68] P. M. McCarthy, “An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld),” Ph.D. dissertation, The University of Memphis, 1993.
- [69] J. Torruella and R. Capsada, “Lexical statistics and tipological structures: A measure of lexical richness,” *Procedia - Social and Behavioral Sciences*, vol. 95, pp. 447 – 454, 2013.
- [70] P. M. McCarthy and S. Jarvis, “Mtd, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment,” *Behavior Research Methods*, vol. 42, no. 2, pp. 381–392, 2010.
- [71] X. Lu, *Computational Methods for Corpus Annotation and Analysis*. Springer Netherlands, 2014.
- [72] IBM Corp., *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, NY: IBM Corp., 2011.
- [73] G. D. Magoulas, M. N. Vrahatis, and G. S. Androulakis, “Improving the convergence of the backpropagation algorithm using learning rate adaptation methods,” *Neural Computation*, vol. 11, no. 7, pp. 1769–1796, 1999.
- [74] L. WEBER and P. KLEIN, *Aplicação da lógica fuzzy em software e hardware*. Editora da ULBRA, 2003.
- [75] G. C. dos Santos, P. S. Thomaz, F. M. Ribeiro, J. F. Araújo, and V. L. D. de Mattos, “Influência do método de defuzzificação em mensurações com controladores fuzzy,” *Blucher Marine Engineering Proceedings*, vol. 1, no. 1, pp. 845–852, 2014.
- [76] X. He and A. Lapedes, “Nonlinear modeling and prediction by successive approximation using radial basis functions,” *Physica D: Nonlinear Phenomena*, vol. 70, no. 3, pp. 289 – 301, 1994.
- [77] A. d. P. Braga, A. P. d. L. F. d. Carvalho, and T. B. Ludermir, *Redes neurais artificiais: teoria e aplicações*. LTC Editora, 2007.
- [78] M. Mak, W. Allen, and G. Sexton, “Comparing multi-layer perceptrons and radial basis functions networks in speaker recognition,” *Journal of Microcomputer Applications*, vol. 16, no. 2, pp. 147–159, 1993.
- [79] S. Chen, C. F. Cowan, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis function networks,” *IEEE Transactions on Neural Networks*, vol. 2, no. 2, pp. 302–309, 1991.
- [80] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1143.
- [81] G. Bugmann, “Normalized gaussian radial basis function networks,” *Neurocomputing*, vol. 20, no. 1, pp. 97–110, 1998.
- [82] F. Heimes and B. van Heuveln, “The normalized radial basis function neural network,” in *IEEE International Conference on Systems, Man, and Cybernetics*, vol. 2, 1998, pp. 1609–1614.
- [83] W. Duch and N. Jankowski, “Survey of neural transfer functions,” *Neural Computing Surveys*, vol. 2, no. 1, pp. 163–212, 1999.
- [84] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014.
- [85] D. M. W. Powers, “Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [86] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton: Butterworth-Heinemann, 1979.
- [87] J. Piskorski and R. Yangarber, “Information extraction: Past, present and future,” in *Multi-source, Multilingual Information Extraction and Summarization*. Springer Berlin Heidelberg, 2013, pp. 23–49.
- [88] F. Pianosi, K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, and T. Wagener, “Sensitivity analysis of environmental models: A systematic review with practical workflow,” *Environmental Modelling & Software*, vol. 79, pp. 214–232, 2016.
- [89] K. Pearson, “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*, vol. 50, no. 302, pp. 157–175, 1900.
- [90] F. Yates, “Contingency tables involving small numbers and the x2 test,” *Supplement to the Journal of the Royal Statistical Society*, vol. 1, no. 2, pp. 217–235, 1934.
- [91] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002.
- [92] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Pearson Education, 2009.
- [93] F. Provost and T. Fawcett, “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*, 1997, pp. 43–48.
- [94] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.
- [95] B. W. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme,” *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.
- [96] R. Likert, “A technique for the measurement of attitudes,” *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.
- [97] F. Komesu and L. Tenani, “Considerações sobre o conceito de “internetês” nos estudos da linguagem,” *Linguagem em (Dis)curso*, vol. 9, pp. 621 – 643, 2009.
- [98] L. V. Avanço, “Sobre normalização e classificação de polaridade de textos opinativos na web,” Master’s thesis, Instituto de Ciências Matemáticas e de Computação (ICMC/USP), 2015.
- [99] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” in *The Handbook of Brain Theory and Neural Networks*, 1998, pp. 255–258.