Um Método de Avaliação da Importância de Medidas para Predição da Nota de um Filme

Gilvan V. Magalhães Junior[1], **Roney L. de S. Santos**[1], **Raimundo S. Moura**[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI)
64.049-550 – Teresina – PI – Brasil

gilvanvmj@gmail.com, roneylira@hotmail.com, rsm@ufpi.edu.br

Abstract. *Analyzing information about movies before watching them is a common practice performed by the audience. But, the large amount of reviews in Online Social Networks and Web sites makes this task almost impossible. As a possible solution to such a problem, some sites use filters based on the scores given by users in their reviews. However, this practice can result in users using the scores as a way to get highlighted. This work proposes an automatic method to analyze the importance of measures from the reviews with the intention of predicting the users' scores. An experiment was carried out with five films of different genres and sets of reviews defined by three thresholds. The experiment showed that the best results were obtained with the threshold of 0.5 and the most important measures were: number of sentences, amount of criticism and correctness of the review.*

Resumo. *Analisar informações sobre filmes antes de assisti-los é uma prática comum realizada pelo público. Entretanto, a grande quantidade de reviews nas Redes Sociais Online (RSO) e em sites da Web torna essa tarefa quase impossível. Como uma possível solução para tal problema, alguns sites utilizam filtros baseados nas notas dadas pelos usuários em seus reviews. No entanto, essa prática pode resultar em usuários utilizando as notas como um artifício para obter destaque. Este trabalho propõe um método automático para analisar a importância de medidas a partir dos reviews com a intenção de predizer a nota dos usuários. Realizou-se um experimento com cinco filmes de gêneros diferentes e conjuntos de reviews definidos por três limiares. O experimento mostrou que os melhores resultados foram obtidos com o limiar 0,5 e as medidas de maior importância foram: quantidade de sentenças, quantidade de críticas e corretude do review.*

1. Introdução

Devido ao avanço da *Web*, pessoas de todo o mundo deixaram de utilizá-la apenas para buscar informações e começaram a compartilhar suas opiniões e experiências por meio de *reviews* on-line. Fontes de pesquisa tais como fóruns, grupos de discussão, blogs, Redes Sociais *Online* (RSO) e *sites* de críticas, além de facilitar o acesso a essas informações contribuem para a formação ou fortalecimento de opiniões. Nas atividades comerciais, quando uma pessoa tem interesse por um produto ou serviço é comum, para tomada de decisão, 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. [de S. Santos et al. 2016].

No entanto, a quantidade de *reviews* publicados diariamente na *Web* é grande e abrange diversos assuntos e opiniões, o que gera a questão sobre quais *reviews* devem ser lidos por exemplo, quando se quer saber sobre a opinião cibernética a respeito de um filme específico. Essas informações originadas de diversas fontes estão distribuídas de forma não estruturada. Com a difusão da *Web*, a grande quantidade de dados faz com que a análise humana se torne uma tarefa impossível, sendo necessária a criação de métodos automáticos para analisar os dados [Liu 2010].

Como uma solução para tal problema, alguns sites utilizam filtros baseados em votos de utilidade de *review* para retornar os melhores, mas tal medida pode não ser a melhor, tendo em vista que novos *reviews* não possuem votos. Outro método utilizado em *sites* é a classificação que o próprio usuário pode dar ao filme por meio de estrelas ou notas, o que pode implicar no problema de usuários classificando filmes com nota máxima apenas com o intuito de ter destaque no *site*. A Figura 1 mostra um *review* de usuário que não assistiu o filme e ainda assim atribuiu uma nota a ele.

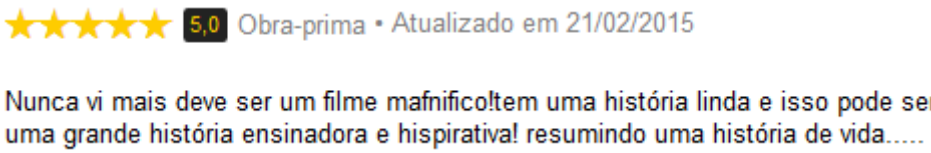

Figura 1. Exemplo de *review*

Este trabalho propõe um método automático de definição da importância das medidas de *reviews* on-line de usuários do *site* Adoro Cinema[1] e predição de suas notas a filmes, utilizando Rede Neural Artificial (RNA). Optou-se por utilizar o Adoro Cinema pelo fato deste possuir o mais abrangente banco de dados sobre cinema no Brasil com mais de 600 mil usuários cadastrados e mais de 80 mil críticas de filmes escritas pelos usuários[2].

O restante deste trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia a ser utilizada. Na Seção 4 são apresentados os resultados e discussões sobre o trabalho. Por fim, a Seção 5 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

[Turney 2002] propôs um algoritmo simples de aprendizado não-supervisionado para classificar comentários como recomendados ou não recomendados. Na prática, o autor avaliou o comentário pela orientação semântica de suas frases contendo adjetivos ou advérbios. O primeiro passo foi identificar as frases que continham adjetivos ou advérbios utilizando um etiquetador. Em seguida, foi analisada a orientação semântica de cada frase extraída (frase positiva ou negativa). Por fim, o terceiro passo foi classificar o *review*

[1]http://www.adorocinema.com
[2]http://www.adorocinema.com/servicos/sobre-nos/

como recomendado ou não recomendado baseado na média da orientação semântica das frases extraídas.

O trabalho de [Barbosa et al. 2016] teve como objetivo estudar formas de quantificar e prever a percepção de utilidade de *reviews* de usuários na *site* Steam. Para isso, eles construíram uma RNA com 11 entradas referentes ao usuário e ao *review*: i) média de votos por *review* do usuário; ii) nota média do usuário; iii) quantidade de horas jogadas; iv) quantidade de amigos; v) padrões linguísticos; vi) legibilidade do *review*; vii) quantidade de palavras; viii) quantidade de sentenças; ix) diferença entre avaliação do usuário e média do produto; x) quantidade de palavras monossílabas; xi) diferença de dias entre o lançamento do produto e postagem. Os resultados deste estudo foram positivos e inferem que as principais características importantes para percepção de utilidade são a reputação do autor e a quantidade de horas jogadas pelo autor.

Em seu trabalho, [Schmit and Wubben 2015] compilaram um corpus de *tweets* para prever os escores de classificação de filmes recém-lançados na IMDb[3]. As previsões foram feitas com diversos algoritmos de aprendizagem de máquina, explorando ambos os métodos de regressão e classificação. Em seus estudos, eles exploraram o uso de vários tipos diferentes de características textuais nas tarefas de aprendizado de máquina. Os resultados mostraram que o desempenho de predição baseado em características textuais derivadas do corpus de *tweets* melhorou na linha de base tanto para regressão como para tarefas de classificação.

[Ganu et al. 2013] propuseram métodos para derivar uma classificação baseada em texto a partir do corpo de *reviews*. Em seguida, agruparam usuários semelhantes em conjunto usando técnicas de *soft clustering* com base nos tópicos e sentimentos que aparecem nos *reviews*. Os resultados do trabalho mostraram que o uso de informações textuais resulta em melhores previsões de nota de *reviews* do que aquelas derivadas das classificações de estrelas numéricas dadas pelos usuários.

A proposta de [Kim et al. 2006] foi um algoritmo para avaliar automaticamente a utilidade e classificar o *review* de acordo com o resultado do algoritmo. Explorando a multiplicidade de avaliações feitas pelos usuários na *Amazon*[4], foi treinado um sistema de regressão SVM (*Support Vector Machine*) para aprender uma função de utilidade e em seguida aplicar a função para classificar os *reviews* sem etiqueta. Também foi realizada uma análise detalhada de diferentes aspectos para estudar a importância de várias classes de aspectos na captura de utilidade, descobrindo assim que as características mais úteis foram o tamanho do *review*, seus unigramas e sua classificação de produto.

Destaca-se que o nosso trabalho utiliza padrões linguísticos para extrair a quantidade de características de um *review*, adaptados da ideia de [Turney 2002]. O método proposto é semelhante ao trabalho de [Barbosa et al. 2016], porém, neste trabalho construímos RNAs com 9 e 7 entradas para definição da importância das medidas de *reviews* e predição da nota de usuários a filmes, explicadas na Seção 4. Diferente do trabalho de [Ganu et al. 2013], nosso trabalho não utilizou sentimentos que aparecem nos *reviews*, no entanto poderiam ter sido explorados.

[3]http://www.imdb.com/

[4]https://www.amazon.com

3. Metodologia

O modelo proposto para analisar a importância de medidas extraídas de *reviews* de usuários do domínio de cinema e fazer a predição das notas dadas aos filmes utiliza uma RNA MLP contendo três camadas, como pode ser visto na Figura 2.

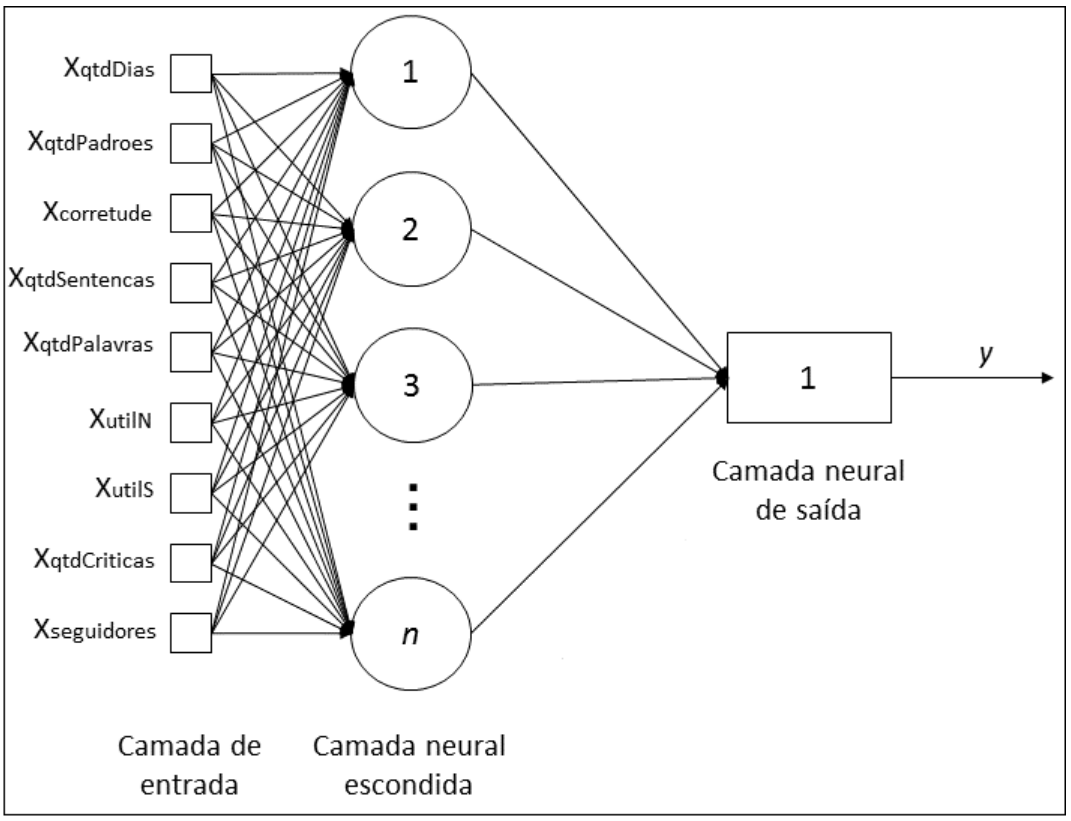

Figura 2. Estrutura geral da Rede Neural Artificial

As seguintes medidas são utilizadas como entradas do modelo: nota em estrelas do filme (valor dado pelo autor entre 0 e 5), número de seguidores do autor ($seguidores$), número de críticas publicadas por ele ($qtdCriticas$), votos de utilidade ($utilN$ e $utilS$), quantidade de palavras ($qtdPalavras$), quantidade de sentenças ($qtdSentencas$), quantidade de palavras corretas ($corretude$), quantidade de características ($qtdPadroes$) e número de dias corridos desde a publicação do *review* ($qtdDias$).

A nota em estrelas, número de seguidores, número de críticas, votos de utilidade e o número de dias corridos são obtidos diretamente do *review*. A quantidade de palavras assim como a quantidade de sentenças podem ser obtidas por meio de expressões regulares. No caso da medida quantidade de sentenças, a cada um ou vários sinais de pontuação (ponto, exclamação ou interrogação) encontrados caracterizam uma sentença, por exemplo: “Amei...”, “O melhor filme do ano, com certeza!!” ou “Bem, por onde começar?”. A medida de quantidade de palavras corretas, pode ser obtida como o uso de um léxico ou dicionário.

A medida referente à quantidade de características pode ser alcançada por meio das classes gramaticais dos tokens, definidas por um etiquetador. As características foram identificadas utilizando reconhecimento dos padrões linguísticos pré-definidos, mos-

trados na Tabela 1, os quais foram adaptados para o Português com base nos padrões linguísticos de [Turney 2002]. A contagem das características se deu pelo número de padrões linguísticos encontrados em cada crítica.

Tabela 1. Padrões linguísticos identificados

Padrão	Padrão linguístico
1	<SUBS> <ADV>? <ADJ>
2	<ADJ> <ADV>? <SUBS>
3	<SUBS> <V> <ADJ>

Para melhor entendimento dos padrões, seguem abaixo alguns exemplos práticos de reconhecimento:

Padrão 1 "filme perfeito", "filme muito bom", "ator fantástico"

Padrão 2 "grandes filmes", "muito bom filme", "melhor livro"

Padrão 3 "filme é ótimo", "filme foi perfeito", "Vorazes aproveitou bem"

3.1. Ajustes da RNA

O modelo proposto utiliza uma RNA *Multilayer Perceptron* (MLP) como aproximador de função para implementação do método por ser um aproximador universal. O teorema de aproximação universal aplicado ao MLP fornece a base necessária para definir as configurações estruturais dessas redes, a fim de mapear funções algébricas [da Silva et al. 2010]. Além disso, o modelo de RNA permite capturar relações não-lineares nos dados sem a necessidade de especificação prévia [Lee and Choeh 2014].

A camada escondida é composta de neurônios e é responsável por todo o processamento interno da rede. A definição da topologia é feita baseada no algoritmo de Seleção de Arquitetura Especializada [IBM Corp. 2011b], que em meio a um intervalo de possíveis valores de neurônios seleciona o "melhor" número de neurônios da camada escondida [de S. Santos 2017]. No entanto, é definido e o algoritmo seleciona o . A camada de saída, constituída de apenas um neurônio, é responsável pela produção do resultado final, que representa a nota predita do *review*. Na camada de saída foi utilizada a função identidade pois o neurônio de saída executa apenas uma combinação linear de funções.

As medidas obtidas de cada *review* devem ser normalizadas e em seguida submetidas como entrada para a RNA. A divisão dos *reviews* em amostras de treino, teste e *holdout* foram especificadas com os valores 9, 1, 0 que representam cada amostra respectivamente e correspondem a 90%, 10% e 0%. O método de particionamento 9, 1, 0 obteve melhor resultados em meio a testes utilizando também os valores 7, 3, 0 (70%, 30% e 0%) e 2, 1, 1 (50%, 25% e 25%), valores empiricamente definidos.

O tipo de treinamento em lotes foi utilizado. Ele atualiza os pesos sinápticos somente após passar todos os registros de dados de treinamento, ou seja, o treinamento em lote usa informações de todos os registros no conjunto de dados de treinamento. Esse tipo de treinamento é frequentemente preferido porque minimiza diretamente o erro total [IBM Corp. 2011a]. Ao final da fase de ajustes, a RNA resulta uma topologia com o valor de neurônios na camada escondida e camada de saída contendo a nota predita. As melhores topologias encontradas serão mostradas na Seção 4.2.

4. Resultados e Discussões

Para avaliar o modelo proposto, realizou-se um experimento com um Córpus de *reviews* de filmes coletados no site Adoro Cinema, descrito a seguir.

4.1. Córpus AdoroCinema

Os dados foram coletados do *site* Adoro Cinema no mês de outubro de 2016 no total de 2.581 *reviews* de usuários e as notas de três veículos da imprensa brasileira que possuiam *reviews* em cada um dos cinco filmes (selecionados aleatoriamente). Os filmes escolhidos possuíam mais comentários e pertencem a gêneros distintos com o objetivo de diversificar o perfil do público. O filme 1 de gênero drama tinha 433 *reviews* de usuários, o filme 2 de gênero romance tinha 442, o filme 3 com gênero ficção científica tinha 499, o filme 4 com gênero ação tinha 577 e o filme 5 com gênero aventura tinha 630. Com finalidade de extrair os dados, a linguagem *Python* combinada com o *Selenium WebDriver*[5] foram utilizados.

As informações dos *reviews* estão dispostas dentre os seguintes aspectos: 1) estrelas (representa a nota ao filme); 2) perfil do usuário (contém características do autor, tais como: nome, quantidade de seguidores e quantidade de críticas); 3) texto (representa a opinião sobre o filme); 4) data (representa quando o *review* foi publicado); e 5) votos de utilidade (representa a aceitação do *review* por outros usuários do *site*). As informações contidas no *review* podem ser vistas na Figura 3

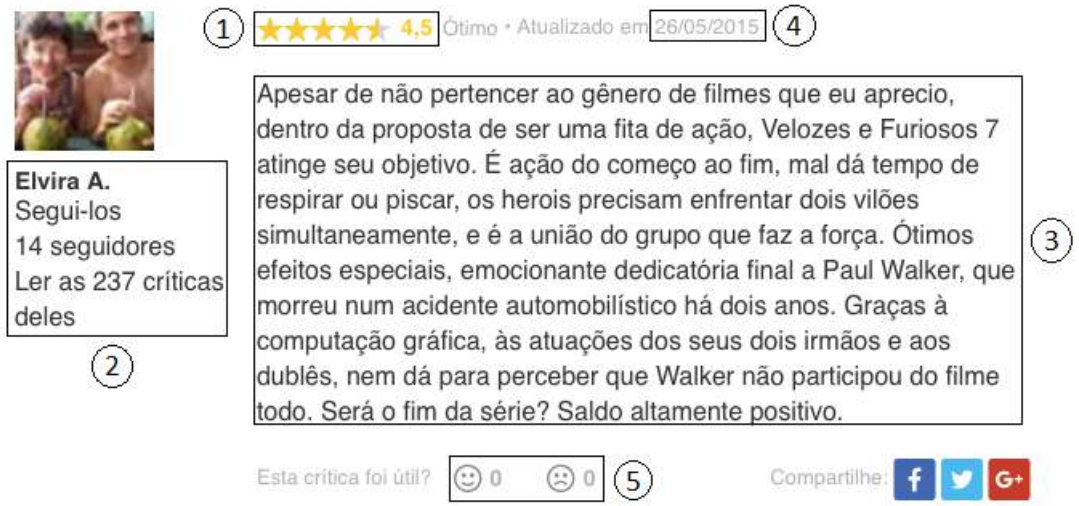

Figura 3. Descrição de informações do *review*

Dentre os *reviews* coletados, foi observado que os votos de utilidade no *site* foram pouco utilizados, ou seja, dos 2.581 apenas 290 *reviews* receberam votos de útil (783 votos) ou não útil (696 votos). Os demais *reviews* possuem o valor zero nestas variáveis. Portanto, para analisar a influência dessas medidas o experimento foi dividido em dois grupos: o grupo 1, onde as medidas de utilidade foram incluídas e o grupo 2, onde essas medidas foram desconsideradas.

[5]http://www.seleniumhq.org/projects/webdriver/

4.2. Experimento

O valor a ser inferido pela RNA é comparado com uma nota juiz, que foi definida como a média aritmética simples das notas dos críticos da imprensa brasileira que possuíam *reviews* nos cinco filmes escolhidos.

O passo seguinte foi agrupar os *reviews* sob diferentes condições com o intuito de perceber a melhor das topologias utilizadas no experimento e obter as notas preditas pela RNA. Os limiares ($\pm$0,5), ($\pm$1,0) e (todos os *reviews*) foram empiricamente definidos e aplicados aos *reviews* de cada filme, buscando dividi-los por meio da nota em estrelas. Os grupos atendem os seguintes intervalos: (nota juiz - 0,5) $<=$ nota em estrelas $<=$ (nota juiz + 0,5), (nota juiz - 1,0) $<=$ nota em estrelas $<=$ (nota juiz + 1,0) e todos os *reviews*. Para melhor entendimento, a Tabela 2 apresenta a aplicação dos limiares para dois filmes.

Tabela 2. Aplicação dos limiares

Filmes	Nota juiz	Limiar	Intervalo
Filme 1	3,67	0,5	3,17 $<=$ nota em estrelas $<=$ 4,17
		1,0	2,67 $<=$ nota em estrelas $<=$ 4,67
		todos	0,0 $<=$ nota em estrelas $<=$ 5,0
Filme 2	3,0	0,5	2,5 $<=$ nota em estrelas $<=$ 3,5
		1,0	2,0 $<=$ nota em estrelas $<=$ 4,0
		todos	0,0 $<=$ nota em estrelas $<=$ 5,0

Para cada filme, foram selecionados os *reviews* com as notas pertencentes ao intervalo dos limiares e, em seguida, submetidos ao modelo da RNA considerando duas variações: i) grupo 1, incluindo as medidas de utilidade ($utilN$ e $utilS$); e ii) grupo 2, sem as medidas de utilidade. Ressalta-se que nossa intenção é obter as notas preditas e analisar as melhores topologias.

Posteriormente, para efetivar a validação dos resultados, foi feita a comparação entre as notas em estrela dos *reviews* coletados e as notas preditas pela RNA por meio do módulo da diferença entre elas. Quando o valor do módulo foi inferior ou igual a 0,5, a nota predita foi considerada correta. Caso contrário, a nota predita foi considerada errada. O valor de aproximação 0,5 foi empiricamente definido baseado na aproximação numérica, mas é importante ressaltar que quanto menor for esse valor, maior será a confiabilidade dos *reviews*. Na Tabela 3 são mostrados os resultados das RNAs.

Observa-se na Tabela 3 que a taxa de acerto é melhor quando as medidas de utilidade são usadas, exceto no filme 2 (limiar 0,5) e no filme 4 (limiar todos). No filme 2 com o limiar 0,5, a RNA pode não ter conseguido generalizar da forma esperada as entradas referentes à utilidade, possivelmente devido a baixa qualidade dos *reviews* desse limiar. No filme 4 com o limiar todos, a exceção ocorreu devido a taxa de *reviews* com votos de utilidade ter sido a menor entre os filmes (8 de 543). Observa-se também que o limiar 0,5 foi melhor para todos os filmes, independente das medidas de utilidade. Isto pode ser justificado pelo fato deste limiar agrupar os usuários que atribuíram notas em estrelas mais próximas dos valores da nota juiz. Adicionalmente, foi possível constatar que quanto mais se amplia o limiar, menos preciso se torna o modelo da RNA.

Tabela 3. Resultados das RNAs

Filme	Limiar	Quantidade de *reviews*	Acerto com utilidade	Acerto sem utilidade
Filme 1 (drama)	0,5	58	100%	100%
	1,0	122	77%	66%
	todos	433	79%	44%
Filme 2 (romance)	0,5	73	75%	82%
	1,0	125	53%	42%
	todos	442	51%	41%
Filme 3 (ficção científica)	0,5	70	96%	84%
	1,0	139	69%	50%
	todos	499	54%	50%
Filme 4 (ação)	0,5	136	84%	82%
	1,0	543	82%	82%
	todos	577	67%	71%
Filme 5 (aventura)	0,5	216	80%	79%
	1,0	596	74%	70%
	todos	630	65%	52%

Com o intuito de analisar qual a importância de cada medida para a predição das notas de usuários a filmes, os 2.581 *reviews* coletados foram utilizados, considerando as mesmas variações de medidas dos grupos 1 e 2, como entradas para a RNA. Vale ressaltar que no grupo 2 não foram utilizadas as medidas de utilidade ($utilN$ e $utilS$). Verificando os resultados, foi possível perceber que a medida de menor importância no grupo 1 foi $qtdSentencas$ (9,2%) e a medida de maior importância foi $utilN$ (100%). No grupo 2, a medida de menor importância foi $qtdPadroes$ (23%) e a medida de maior importância foi $qtdSentencas$ (100%). Nota-se que $qtdSentencas$ foi a medida de menor importância no grupo 1 e a medida de maior importância no grupo 2, isto pode ser justificado por uma possível pertubação causada pelas medidas de utilidade no modelo. A Figura 4 ilustra os resultados dos grupos 1 e 2.

Por fim, analisando os resultados das RNAs apenas para o grupo 2 e considerando os *reviews* de cada filme individualmente, foi possível também identificar as melhores topologias para cada limiar, a saber: i)limiar 0,5: a melhor topologia encontrada foi para o filme 1 utilizando 4 neurônios na camada escondida; ii) limiar 1,0: a melhor topologia encontrada foi para o filme 4 utilizando 1 neurônio na camada escondida; iii) limiar todos: a melhor topologia encontrada foi também para o filme 4 utilizando 1 neurônio na camada escondida. Os valores para os três limiares de cada filme não serão apresentados neste artigo por questão de espaço. Destaca-se ainda que a melhor topologia encontrada para a RNA usada na definição da importância das medidas foi com 2 neurônios na camada escondida.

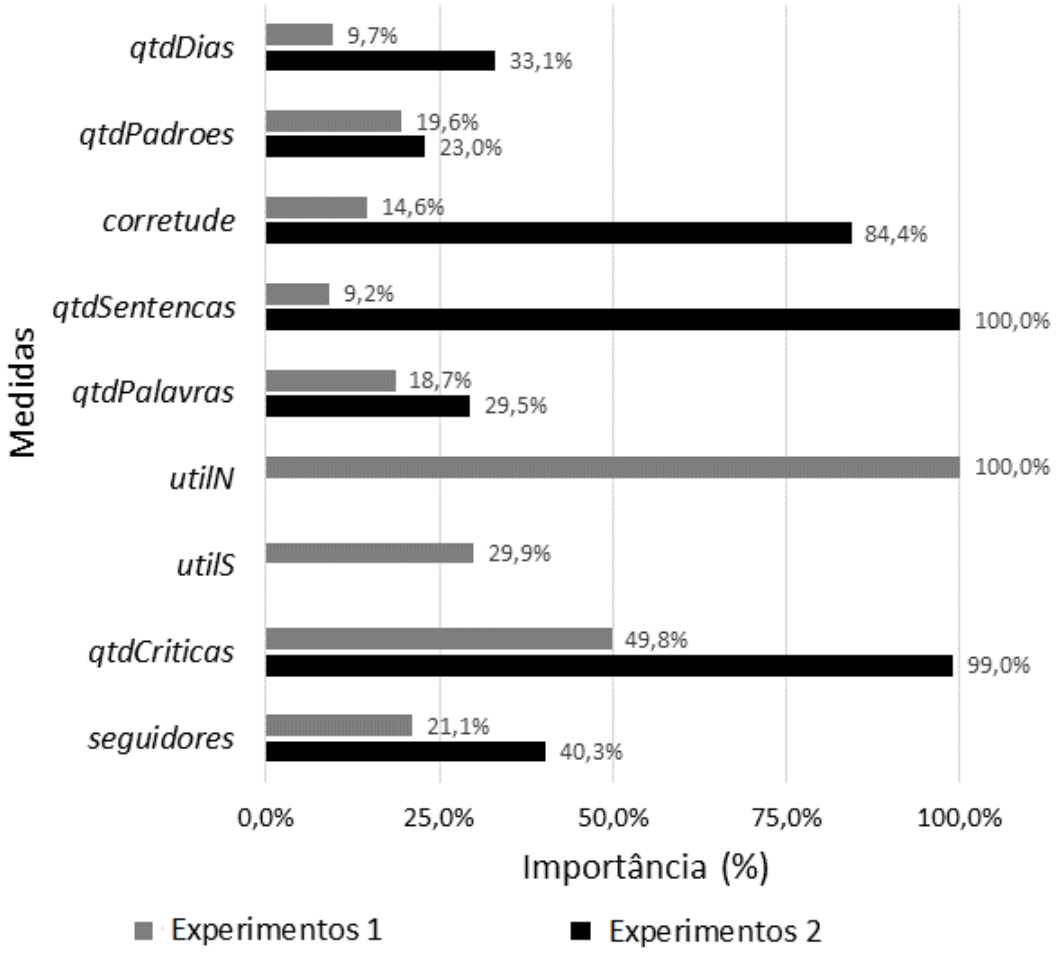

Figura 4. Importância das medidas dos *reviews*

5. Conclusão

Este trabalho apresentou um método automático para definir a importância das medidas obtidas a partir dos *reviews* para predição de notas de usuários aos filmes. Nossos resultados sugerem que as medidas de maior importância encontradas foram quantidade de sentenças, quantidade de críticas e a corretude. Observando os resultados obtidos das RNAs, foi possível perceber que o limiar 0,5 obteve melhor precisão entre os três limiares usados devido a maior proximidade de notas em estrelas dos *reviews* com a nota juiz.

Para trabalhos futuros, pretende-se: i)analisar o impacto dos padrões linguísticos usados no processo de classificação; ii) Adicionar mais palavras ao léxico, como neologismos e gírias comuns, a fim de melhorar a corretude das críticas, por exemplo, "filmaço" e "vilanesca"; iii) Tratar erros provocados por letras repetidas em palavras; e iv) Aplicar o método em outros domínios tais como: músicas, livros, dentre outros.

Referências

Barbosa, J. L. N., Moura, R. S., and de S. Santos, R. L. (2016). Predicting portuguese steam review helpfulness using artificial neural networks. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web*, pages 287–293.

da Silva, I. N., Spatti, D. H., and Flauzino, R. A. (2010). *Redes Neurais Artificiais: para engenharia e ciencias aplicadas*. Artliber.

de S. Santos, R. L. (2017). Um estudo comparativo entre abordagens baseadas em sistemas fuzzy e redes neurais artificiais para estimar a importância de comentários sobre produtos e serviços. Master's thesis, Universidade Federal do Piauí, Teresina, PI.

de S. Santos, R. L., Vieira, J. P. A., Barbosa, J. L. N., Sá, C. A., Moura, E. G., Moura, R. S., and de Sousa, R. F. (2016). 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*, pages 153–163.

Ganu, G., Kakodkar, Y., and Marian, A. (2013). Improving the quality of predictions using textual information in online user reviews. *Information Systems*, 38(1):1–15.

IBM Corp. (2011a). IBM Knowledge Center Training (Multilayer Perceptron).

IBM Corp. (2011b). *IBM SPSS Statistics for Windows, Version 20.0*. IBM Corp., Armonk, NY.

Kim, S.-M., Pantel, P., Chklovski, T., and Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430.

Lee, S. and Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Syst.Appl.*, 41(6):3041–3046.

Liu, B. (2010). *Sentiment Analysis and Subjectivity*. Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group.

Schmit, W. and Wubben, S. (2015). Predicting ratings for new movie releases from twitter content. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2015, 17 September 2015, Lisbon, Portugal*, pages 122–126.

Turney, P. D. (2002). 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*, pages 417–424.