Submission date: 13/09/2018
Resubmission date: 22/04/2019
Camera ready submission: 03/07/2019

1st round notification: 25/03/2019
2nd round notification: 24/06/2019
Available online: 18/07/2019
Publication date: 23/08/2019

Section: extended version
Extended versions of papers presented at the VII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2018)

Estudo sobre Métricas para Definir Reputação do Autor de Comentários em Sites de Vendas de Produtos

Title: A Study about Metrics for Defining the Author Reputation of Web Comments on Products

Carlos Augusto de Sá[1], Raimundo Santos Moura[2]

[1] Colégio Técnico de Teresina – Universidade Federal do Piauí (UFPI)
Teresina, Piauí – Brasil
[2] Departamento de Computação – Universidade Federal do Piauí (UFPI)
Teresina, Piauí – Brasil

carlos.sa@ufpi.edu.br, rsm@ufpi.edu.br

Abstract. *Knowing the reputation of the author of opinion texts on the Web is of utmost importance for the development of systems based on open data. This paper presents a study on measures used in the process of evaluating the author's reputation on product sales sites. Two experiments were carried out with neural networks Multilayer Perceptron (MLP) and Radial Basis Function (RBF), and the results show that the MLP gave slightly better performance, but not significantly so. In addition, an experiment was carried out to compare the TOP(X) approach, which is used to infer the best comments, with the new approach that uses MLP in the author's reputation dimension. The results showed that the new approach obtained a gain in the classification of the importance of the comments. In addition, a fourth experiment with other machine learning algorithms was performed to observe the behavior of the data.*

Keywords. *Author Reputation; Opinion mining, Artificial neural networks.*

Resumo. *Conhecer a reputação do autor de textos opinativos na Web é de suma importância para o desenvolvimento de sistemas baseados em dados abertos. Este artigo apresenta um estudo sobre medidas usadas no processo de avaliação da reputação do autor em sites de vendas de produtos. Realizou-se dois experimentos com as redes neurais Multilayer Perceptron (MLP) e Radial Basis Function (RBF), sendo que a rede MLP obteve melhor desempenho. Em um terceiro experimento, comparou-se a abordagem TOP(X) original, usada para inferir os melhores comentários, com um novo modelo que utiliza rede MLP na dimensão da reputação do autor. Considerando os comentários excelentes e bons, a nova abordagem apresentou resultados significativamente superiores. Adicionalmente, foi realizado um quarto experimento com outros algoritmos de aprendizagem de máquina (AM) para observar o comportamento dos dados.*

Palavras-Chave. *Reputação do autor; Mineração de opinião; Redes neurais artificiais.*

Cite as: *Sá, C. A.& Moura, R. S. (2019). A Study about Metrics for Defining the Author Reputation of Web Comments on Products (Estudo sobre Métricas para Definir Reputação do Autor de Comentários em Sites de Vendas de Produtos). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 12(3), 06-23.*

1. Introdução

De acordo com os dados do estudo *Global Digital Report 2018*[1], realizado pelas empresas *We Are Social* e *Hootsuite*, temos 7,5 bilhões de pessoas no mundo, sendo que mais da metade acessa à Internet e usa um aparelho celular. O mesmo estudo aponta para 3,1 bilhões de usuários ativos nas diversas mídias sociais e para o aumento constante do uso de redes sociais via dispositivos móveis. Nos últimos anos o comportamento desses usuários vem mudando, pois além de consumir conteúdos, eles também estão expondo suas opiniões e experiências, seja sobre um produto que adquiriram, um local que visitaram ou um serviço que utilizaram, proporcionando assim uma maior interação. Atualmente, com o crescimento do volume de informações disponíveis e com o avanço da Computação, a área de Processamento de Linguagem Natural (PLN) ganhou bastante destaque por realizar análises de dados de maneira mais eficiente.

Análise de Sentimentos ou Mineração de Opinião é uma subárea de PLN que envolve Ciência da Computação, Linguística e Inteligência Artificial e tem atacado o problema de manipular grandes volumes de dados através de técnicas que analisam a linguagem escrita ou falada [Jackson and Moulinier 2007]. Um desafio da área se encontra na filtragem ou pré-processamento de comentários Web, já que existe uma tendência dos usuários escreverem usando muitas gírias, o que dificulta o trabalho das ferramentas tradicionais de PLN. Outro detalhe a ser considerado é a quantidade de *spam*, textos de baixa qualidade, erros ortográficos, *emoticons*, "internetês"[2] e informações falsas [Liu 2011]. Além dessas considerações, às vezes os comentários apresentam sarcasmos e ironias, que são difíceis de serem captados pelas técnicas atuais de PLN. No entanto, existem esforços no sentido de resolver esses problemas, como os trabalhos de [Hartmann et al. 2014, Carvalho et al. 2009, Goncalves et al. 2015].

Uma característica do ser humano é frequentemente buscar opiniões de outras pessoas sobre um produto ou serviço antes de adquiri-lo. No ambiente Web, destaca-se que avaliações positivas a respeito de um produto ou serviço trazem ao novo consumidor mais segurança no processo de compra, porém, avaliações negativas também podem auxiliar na escolha, gerando um incremento nas vendas [Hamilton et al. 2014]. Geralmente, comentários negativos são escritos de forma mais crítica e apresentam boa legibilidade, sendo, algumas vezes, melhores do que comentários positivos. A Figura 1 apresenta um comentário negativo retirado do site da empresa Amazon, enfatizando o recurso conhecido como *utilidade do review*. Com este recurso, o usuário pode, ao terminar de ler, marcar a opção indicando se o comentário foi útil para ele. Desta maneira, quanto mais votos "Sim" um comentário possuir, melhor classificado ele será. Porém, uma desvantagem dessa medida é que comentários recentes e com alta significância ao consumidor, são ignorados por terem poucos votos [Li et al. 2013].

[1] https://digitalreport.wearesocial.com/

[2] Internetês á um neologismo (palavra: Internet + sufixo: ês) que designa a linguagem utilizada no meio virtual.

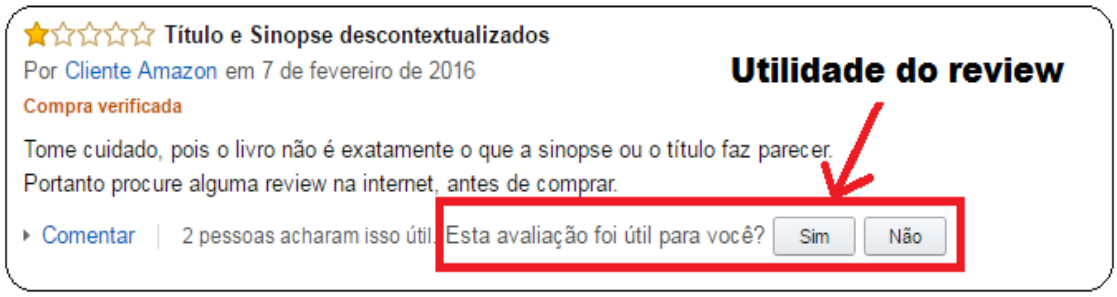


Figura 1. Review negativo do site Amazon

Um problema que os sites de *e-commerce* apresentam é o fato de possuírem cadastros independentes para os seus usuários e o acesso a esses dados não ser permitido, dificultando a coleta de informações. Desta forma, esta pesquisa sugere que os sites de *e-commerce* possam rever as suas políticas de privacidade no futuro. Uma tentativa de solucionar o entrave é ligar o perfil dos usuários com a suas contas em RSOs populares como *Twitter* e *Facebook*. Nas RSOs é possível explorar os relacionamentos de amizades, seguidores e curtidas para analisar a repercussão de um dado comentário. Destaca-se que as empresas tentam explorar ao máximo esse novo tipo *marketing*, o que se confirma com a grande quantidade de perfis nesta rede social.

Com o objetivo de identificar os comentários mais relevantes, em [De Sousa et al. 2015] os autores propuseram uma abordagem para inferir os melhores comentários sobre produtos ou serviços, denominada TOP(X), que utiliza um *Sistema Fuzzy* com três variáveis de entrada: reputação do autor, número de tuplas <característica, palavra opinativa>, e riqueza de vocabulário; e uma variável de saída: grau de importância do comentário, representado pela variável "k", (ver Figura 2). No entanto, os autores utilizam somente a quantidade de comentários publicados para avaliar a reputação do autor, o que pode ser facilmente questionável, por exemplo, um ***spammer***[3] é considerado um bom autor.

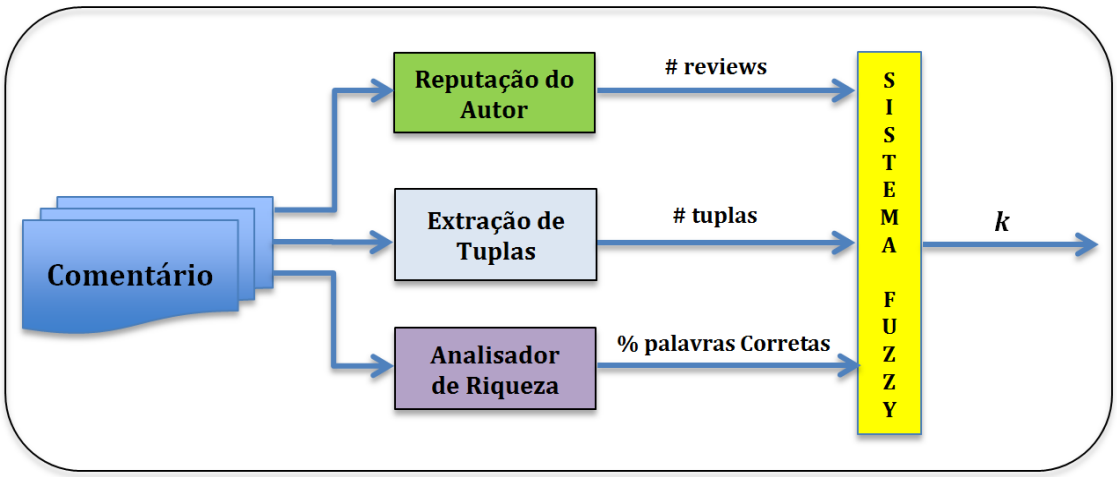


Figura 2. Abordagem TOP(X) proposta por [De Sousa et al. 2015]

[3] *Spammer* é um usuário que posta muitas propagandas sem a permissão dos demais usuários.

Com a intenção de explorar o impacto da variável reputação do autor na abordagem TOP(X), este artigo apresenta um estudo utilizando Rede Neural Artificial (RNA) para analisar um conjunto de medidas referentes ao autor do comentário e definir quais são as mais relevantes no processo de avaliação. Outra contribuição é fazer uma comparação entre a abordagem TOP(X) original com a nova abordagem que utiliza uma RNA na dimensão da reputação do autor. A questão de pesquisa que queremos responder é se o uso de RNA para inferir a reputação do autor de comentários Web melhora a acurácia do modelo de ranqueamento de comentários.

É importante mencionar que este artigo trata-se de uma versão estendida de trabalho previamente apresentado pelos autores em um evento científico. A extensão está focada em dois aspectos: i) análise dos dados usando outros algoritmos de aprendizagem de máquina para a tarefa de regressão; e ii) análise adicional de erros e discussão sobre ameaças à validade dos experimentos realizados.

O restante deste artigo está organizado de seguinte maneira: na Seção 2 apresenta-se alguns trabalhos sobre reputação de autor em ambientes Wiki e em RSOs. Na Seção 3 descreve-se a abordagem proposta para analisar o conjunto de medidas sobre reputação do autor, usando RNAs. A Seção 4 destaca a coleta e a preparação do *Corpus* utilizado nos experimentos. A Seção 5 descreve os experimentos realizados com as RNAs e com os algoritmos de regressão e discute os resultados obtidos. Na Seção 6 apresenta-se a análise de erros e uma discussão sobre ameaças à validade dos experimentos. Por fim, a Seção 7 destaca as principais contribuições e trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, não existe uma definição formal para reputação de autor de comentários Web. Porém, algumas medidas têm sido propostas em sistemas de avaliação. Segundo Li et al. (2014), a reputação de um usuário em um sistema de avaliação pode ser medida de acordo com as informações postadas por tal usuário, desta forma, quanto mais avaliações justas e confiáveis forem escritas, melhor a reputação. Adicionalmente, os autores alertam para a importância das abordagens minimizarem, ao máximo, a atuação dos *spammers*, que são usuários que postam propagandas sem a permissão dos demais usuários.

Para Jones, Hesterly and Borgatti (1997), a reputação envolve uma estimativa do caráter, habilidades e confiabilidade de um indivíduo. Segundo os autores, a reputação reduz o comportamento de incerteza por prover informações a respeito da confiabilidade e boa vontade dos outros. Na área de governança corporativa, uma boa reputação traz consequências econômicas para as empresas.

Diversos autores têm investigado sobre avaliação de reputação de autor na Web e nas redes sociais, com destaque para os ambientes *Wiki* e para o *Twitter*. Os ambientes *Wiki* se caracterizam por permitir a colaboração mútua entre os usuários na produção de artigos dos mais variados temas. Um problema inerente à esta liberdade é a possibilidade de se ter artigos de baixa qualidade, especialmente pela atuação de vândalos[4]. As principais formas de avaliar a reputação do autor nos ambientes *Wiki* são:

[4] Vândalos são usuários que editam os artigos com informações fora do contexto.

- **Histórico das edições**: os autores utilizam o histórico das páginas em busca de padrões de *edits* por parte dos usuários. Wöhner et al. (2011) destacam que as contribuições persistentes de usuários na Wikipedia duram em média 14 dias sem sofrer modificações. Eles classificam os usuários como autores *vândalos* ou *regulares*. Halfaker et al. (2009) definem um juiz para classificar um artigo como *aceito* ou *rejeitado* pela comunidade *Wiki* baseado em três características: qualidade dos colaboradores, experiência e no conteúdo postado. Adler and De Alfaro (2007) indicam que autores *Wiki* ganham reputação quando seus *edits* são preservados por autores subsequentes e perdem reputação quando seus *edits* são desfeitos em um período curto de tempo. Adler e seus amigos definiram, então, o sistema de reputação *WikiTrust* [Adler et al. 2010] baseado em três características: qualidade do *edit*, reputação do autor e reputação do conteúdo.
- **Contexto social**: Zhao et al. (2010) definiram a *SocialWiki*, um protótipo de sistema *Wiki*, que aproveita o poder das redes sociais para gerenciar automaticamente reputação e confiança para os usuários *Wiki*, baseado no conteúdo que eles contribuem e nas avaliações que eles recebem de outros usuários. Os autores consideram como colaboradores de um artigo, os usuários com interesses em comum, porém eles não descreveram a fórmula para calcular a reputação.
- **Mecanismos de recompensa**: Hoisl et al. (2007) focaram sobre mecanismos de recompensa social, tais como aceitação, poder e *status*, para ranquear autores que mais colaboram com boas contribuições. Os autores concluíram que a abordagem de recompensa baseada em motivação pode produzir artigos de alta qualidade.

É importante destacar que a contribuição dos trabalhos que exploram o ambiente Wiki está na persistência das colaborações, ou seja, quanto mais tempo um *edit* persistir, melhor a reputação do autor.

Com relação a rede social *Twitter*, destacam-se os trabalhos para identificar os usuários mais influentes, usuários suspeitos e *spammers*. Kwak et al. (2010) utilizam os dados coletados nos "*trending topics*" (assuntos do momento, em tradução livre) para criar *ranking* dos usuários de acordo com o número de seguidores e o algoritmo de *Page-Rank*. Eles notaram que esses dois *rankings* são similares. Os autores criaram um terceiro *ranking* baseado nos *retweets*, que é o processo de propagar na rede o *tweet* de outro usuário. Eles concluíram que um *retweet* possui alcance de, no mínimo, 1000 usuários, devido à forma de propagação instantânea e que mais de 85% dos tópicos classificados se referem a manchetes de provedores de conteúdo.

Weitzel et al. (2014) definiram medidas baseadas nos *retweets* para calcular a reputação dentro do *Twitter*, abordando informações no domínio da medicina. Os autores concluíram que a maioria dos perfis no *Twitter* são individuais ou de *blogs* e que a aplicação das medidas baseadas em *retweets* conseguem identificar os usuários mais populares dentro da rede.

Weng et al. (2010) propuseram a medida *TwitterRank*, baseada no número de seguidores e seguidos do usuário. De acordo com a abordagem dos autores, dados três usuários A, B e C, sendo que C segue A e B; se A e B publicam, respectivamente, 500 e

1.000 *tweets* sobre um dado tópico, então, a influência que B exerce sobre C é duas vezes maior que a influência de A. Ainda sobre medidas de ranqueamento, Cappelletti and Sastry (2012) criaram o algoritmo *IARank*, que observa o potencial de um usuário ampliar uma informação dentro do *Twitter*. Eles consideram a tendência de um usuário ser retuitado ou mencionado e o tamanho da audiência desses retuites ou menções.

No trabalho desenvolvido por Aggarwal and Kumaraguru (2014), os autores identificaram um "mercado negro" que vende/compra contas fraudulentas, curtidas no *Facebook* e até mesmo seguidores no *Twitter* para, artificialmente, melhorarem a reputação social dos usuários. Os autores relatam uma precisão de 88,2% no mecanismo de aprendizagem de máquina supervisionado usado para predizer seguidores suspeitos.

No que se refere a detecção de *spammers*, Wang (2010) definiu reputação do autor como sendo uma relação entre o número de amigos e o número de seguidores. Os resultados obtidos demonstram que o sistema de Wang consegue detectar comportamentos anormais de usuários.

Por fim, no contexto dos *sites* de *e-commerce*, existem soluções que criam *rankings* e filtros dos comentários sobre os produtos para auxiliar os consumidores no momento da compra. Os *rankings* podem ser ordenados por data ou número de estrelas. Adicionalmente, podem existir filtros para listar apenas os comentários positivos, negativos, de compradores verificados ou de produtos de uma determinada característica, por exemplo, produto da cor azul. As Tabelas 1 e 2 mostram o resumo dos trabalhos no ambiente *Wiki* e em redes sociais, respectivamente, que serviram como base para a realização de nossa proposta, detalhando as abordagens e as formas usadas para avaliar a reputação.

Tabela 1. Resumo dos trabalhos relacionados: ambiente *Wiki*

Trabalho	**Descrição**	**Ambiente e Forma de Avaliação**
Wöhner et al. 2011	Utiliza histórico das edições para descobrir que contribuições persistentes duram em média 14 dias.	*Wiki*, Histórico dos *edits*
Halfaker et al. 2009	Utilizam um juiz para classificar um artigo baseado na qualidade das colaborações, experiência e conteúdo postado.	*Wiki*, Histórico dos *edits*
Adler e De Alfaro, 2007 Adler et al. 2010	Definiram o sistema *WikiTrust* baseado na qualidade do *edit*, reputação do autor e reputação do conteúdo. Eles indicam que um autor ganha reputação quando os *edits* são preservados e perde reputação quando o *edit* é desfeito.	*Wiki*, Histórico dos *edits*
Zhao et al. 2010	Deifniram o *SocialWiki* baseado no conteúdo das colaborações e nas avaliações que os usuários recebem.	*Wiki*, Conteúdo social
Hoisl et al. 2007	Criaram um *ranking* dos autores, baseado em mecanismos de recompensa social, tais como aceitação, poder e *status*.	*Wiki*, Mecanismo de recompensa

Tabela 2. Resumo dos trabalhos relacionados: *Redes Sociais*

Trabalho	Descrição	Ambiente e Forma de Avaliação
Kwak et al. 2010	Criaram *rankings* de usuários, baseado no número de seguidores, no algoritmo de *page-rank* e em *retweets*. Eles concluíram que um *retweet* possui alcance de 1000 usuários e que mais de 85% dos tópicos se referem a manchetes de provedores.	*Twitter*, número de seguidores e *retweets*
Weitzel et al. 2014	Definiram medidas baseadas nos *retweets* para calcular a reputação, abordando informações no domínio da medicina. Concluíram que a maioria dos perfis são individuais ou de *blogs*.	*Twitter*, *retweets*
Weng et al. 2010	Propuseram o *TwitterRank* baseado no número de seguidores e seguidos de um usuário.	*Twitter*, número de seguidores e seguidos
Cappelletti e Sastry, 2012	Criaram o IARank para observar o potencial de um usuário em ampliar uma informação. Eles consideram a tendência de um usuário ser retuitado ou mencionado e o tamanho da audiência.	*Twitter*, *retweets* e menções
Aggarwal e Kumaraguru, 2014	Usaram AM supervisionada para predizer usuários suspeitos, que vendem/compram contas fraudulentas, curtidas no *Facebook* e seguidores no *Twitter*.	*Twitter e Facebook,* curtidas e seguidores
Wang, 2010	Para detectar *spammers*, ele definiu reputação do autor como uma relação entre o número de amigos e o número de seguidores.	RSOs, número de amigos e seguidores

Destaca-se que o sistema proposto por Wang (2010) pode ser aplicado a qualquer rede social. No entanto, à luz de nosso conhecimento, não existe nenhuma proposta na literatura para criar um *ranking* de comentários sobre produtos e serviços, analisando as características textuais dos relatos. Assim, nosso artigo investiga se a reputação do autor pode ser calculada a partir de seis medidas, incluindo a descrição textual, que serão abordadas com mais detalhes na próxima seção.

3. Abordagem Proposta

A abordagem proposta neste trabalho visa analisar um conjunto de medidas para definir a reputação do autor de comentários em sites de vendas de produtos. De forma geral, a proposta representa uma adaptação da abordagem Top(X) original, com ênfase na dimensão reputação do autor. O estudo foi conduzido através da aplicação de redes neurais artificiais para inferir a reputação dos autores dos comentários e descobrir a importância de cada medida da entrada.

A Figura 3 mostra uma visão parcial do modelo, considerando apenas a dimensão reputação do autor. Na figura, a variável ‘a’ representa a saída da RNA e indica a reputação do autor normalizada para o intervalo de 0 a 10.

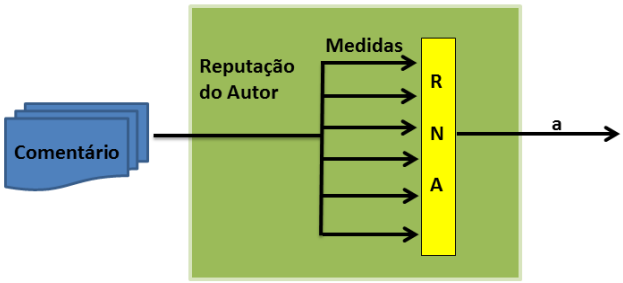


Figura 3. Abordagem proposta

Considerando sites de *e-commerce*, definiu-se seis medidas para avaliar a reputação do autor dos comentários de produtos. Tais medidas foram extraídas levando em conta as informações disponíveis em sites de lojas virtuais e comparadores de preços. Destaca-se que nesta pesquisa outras medidas utilizadas em ambientes Wiki, RSOs, Fóruns e Blogs foram examinadas, por exemplo, número de seguidores e quantidade de edições no texto. No entanto, essas informações não estão disponíveis em sites de *e-commerce*. Assim, sugere-se, então, que esses sites devam expandir suas funcionalidades no sentido de reforçar a importância dos autores e seus relacionamentos. Uma maneira de realizar essa expansão é permitir a integração dos perfis de usuários dos sites com os seus respectivos perfis em redes sociais.

O modelo de RNA proposto usa seis variáveis de entrada, definidas como:

- **DataReview**: a data de escrita do comentário, convertida para dias em comparação com a data inicial de coleta do *Corpus*. Esta informação é importante pois quanto mais recente, mais atualizado o comentário e, hipoteticamente, deveria ser melhor avaliado. No entanto, os comentários que são muito recentes podem ser prejudicados no processo de avaliação geral por não ter tempo hábil para leitura pelos consumidores;
- **DataCadastro**: a data em que o autor fez o seu cadastro no site, convertida para dias em comparação com a data inicial da coleta do *Corpus*. Esta informação é importante pois imagina-se que a reputação de autores experientes seja melhor do que de autores novatos;
- **VotosPositivos**: quantidade de votos positivos atribuídos por outros usuários. A hipótese é que quanto mais votos positivos um autor receber de outros usuários, melhor será a sua reputação;
- **VotosNegativos**: quantidade de votos negativos atribuídos por outros usuários. A importância dos votos negativos é inversamente proporcional aos votos positivos, pois quanto mais votos negativos o autor receber em seus comentários, pior a sua reputação;
- **TotalVotos**: soma dos votos recebidos pelo comentário. De forma geral, imagina-se que quanto mais votos o usuário tenha em seus comentários, sejam positivos ou negativos, melhor a sua reputação pois o mesmo está sendo observado;
- **TotalReviewsAutor**: quantidade de comentários que o autor realizou no site. Esta informação é relevante pois indica a participação ativa do usuário dentro do ambiente.

4. *Corpus*: Coleta e Preparação

Para avaliação da abordagem proposta, foi criado um *Corpus* com 2.433 comentários do site do Buscapé[5], que está disponível em https://goo.gl/g5nrwJ. A decisão de trabalhar com esses comentários deu-se por três razões principais: i) ser o maior site comparador de preços da América Latina; ii) necessidade posterior de comparar o modelo proposto com a abordagem Top(x), que utiliza comentários sobre *smartphones* do referido site; e iii) os comentários serem disponíveis publicamente para coleta com rastreadores Web. É importante destacar que foram usados apenas comentários do site do Buscapé, porém comentários de outros sites de vendas e de outros produtos podem ser analisados sem inviabilizar a abordagem proposta.

Os dados foram coletados nos dias 28 e 29 de setembro de 2016 e são referentes a comentários escritos em português sobre os diversos modelos de *smartphones* existentes no site do Buscapé. Após a exclusão de comentários duplicados e vazios, definiu-se uma amostra aleatória de 2.000 comentários, sendo 1.000 de orientação positiva e 1.000 negativa. Destaca-se que no site do Buscapé, a orientação do comentário é definida pelo próprio autor e, são apresentados aos usuários em guias/abas separadas. No entanto, em uma análise mais detalhada, verificou-se que alguns comentários marcados como positivos eram, na verdade, negativos e vice-versa. Além disso, muitos comentários são considerados neutros.

Para solucionar esse problema, decidiu-se fazer uma revisão manual do *Corpus* quanto à orientação semântica. Ao final do processo, o *Corpus* anotado ficou com 923 comentários positivos, 602 comentários negativos, 141 comentários neutros e 334 considerados "lixo", que são comentários de usuários que declararam "não possuir o produto" e comentários totalmente sem sentido. Os comentários "lixo" foram desconsiderados nas nossas avaliações.

Em um segundo momento, criou-se um *Subcorpus* anotado com a reputação do autor, analisando uma amostra de 323 comentários, considerando o nível de confiança de 95% e margem de erro de 5%. Adicionou-se 33 comentários ao *Subcorpus*, que corresponde a 10% da amostra, totalizando 356, sendo 132 positivos, 131 negativos e 93 neutros, identificado como ***Subcorpus I***. A anotação foi realizada por três alunos de pós-graduação em Ciência da Computação que atuam na área de PLN e que são usuários comuns de *smartphones*. Eles consideraram informações referentes ao autor, como a quantidade de votos positivos em seus comentários, quantidade de votos negativos, total de votos, entre outras medidas. Em seguida, aplicou-se uma nota de 0 a 10 para cada um dos autores dos comentários dentro da amostra definida, sendo guiado unicamente pelas variáveis de entrada.

A Tabela 3 mostra o resultado da avaliação dos especialistas para a reputação dos autores dos comentários. As onze notas atribuídas aos autores foram generalizadas para o universo completo dos comentários através de uma RNA, como será descrito na próxima seção. Esta generalização se dá pela rede neural que infere, a partir das medidas de entrada, a reputação do autor para qualquer comentário dentro do *Corpus*.

[5] http://www.buscape.com.br/

Tabela 3. Resultado da anotação do *Subcorpus I* por especialistas

Reputação	#Total	Reputação	#Total	Reputação	#Total	Reputação	#Total
0	68	3	23	6	9	9	5
1	163	4	11	7	5	10	5
2	50	5	16	8	1		

Para avaliar a acurácia do modelo original definido por De Sousa (2015) e do modelo proposto neste trabalho, foi necessário criar um *Subcorpus* com a avaliação da importância dos comentários, identificado como ***Subcorpus II***. Este *subcorpus* foi definido com 271 comentários selecionados aleatoriamente, sendo 100 positivos, 100 negativos e 71 neutros, considerando a distribuição estatística das orientações positivas, negativas e neutras do *Corpus* principal, sem considerar os comentários do tipo "lixo".

A Tabela 4 apresenta o resultado da avaliação realizada por apenas um especialista (aluno de pós-graduação em Ciência da Computação) quanto à importância dos comentários. Ele levou em consideração três fatores: quantidade de informações destacadas no comentário, riqueza do vocabulário e reputação do autor (nota de 0 a 10), inferida pelo processo de generalização da RNA.

Tabela 4. Resultado da anotação do *Subcorpus II* por especialista

Importância	Quantidade
Excelente (EX)	17
Bom (BM)	24
Suficiente (SF)	145
Insuficiente (IF)	85

Por fim, é importante mencionar que as anotações manuais dos *Corpora* foram realizadas porque não existem recursos linguísticos para a língua portuguesa disponíveis publicamente. Sabe-se também que o processo de anotação manual pode causar um viés nas avaliações e, possivelmente, comprometer a viabilidade da solução. Porém, os riscos da anotação foram minimizados com o envolvimento de especialistas da área de linguística computacional.

5. Experimentos

Para avaliar o modelo de RNA proposto, realizou-se dois experimentos com as arquiteturas *Multilayer Perceptron* (MLP) e *Radial Basis Functions* (RBF), considerando o ***Subcorpus I***. As seis medidas discutidas na Seção 3 foram utilizadas como os neurônios da camada de entrada. Como o objetivo de comparar a abordagem TOP(X) original com a nova abordagem que usa RNA na dimensão da reputação do autor, fez-se um terceiro experimento discutido na subseção 5.3. Por fim, para analisar o comportamento dos dados sobre comentários de *smartphones*, realizou-se um quarto experimento discutido na subseção 5.4, usando outros algoritmos de AM.

5.1. Experimento 1: RNA MLP

No primeiro experimento usamos uma RNA MLP e o melhor ajuste se deu com 8 neurônios na camada escondida e função de ativação *Tangente Hiperbólica*. Na camada de saída utilizamos o atributo de supervisão "ReputacaoManual" como variável dependente para testar a rede, classificando as 11 notas possíveis dos autores (intervalo de 0 a 10) e função de ativação *Softmax*. É importante relatar que o ajuste foi realizado pela ferramenta de análises estatísticas SPSS [SPSS, 2007], através de diversos testes internos para atingir a melhor configuração possível da rede.

A Figura 4 mostra a disposição dos neurônios em cada camada na melhor topologia definida pela execução da RNA MLP. O processo de treinamento e teste foi aplicado com o ***Subcorpus I*** utilizando o método de validação cruzada *10-fold cross validation*.

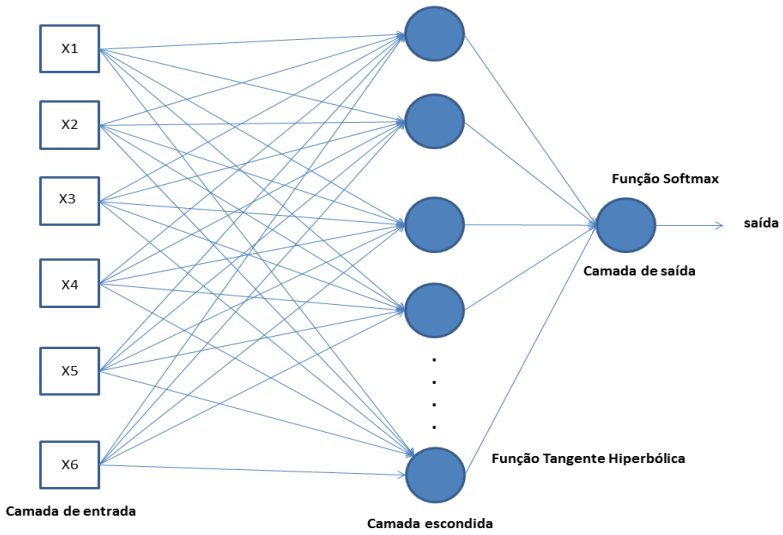


Figura 4. Topologia da RNA MLP

Em redes neurais, o valor da importância de uma variável de entrada é calculado levando em consideração o peso das conexões dos neurônios entre as camadas da rede. Já a importância normalizada é simplesmente os valores de importância divididos pelos maiores valores de importância e expressos em porcentagens.

A Tabela 5 apresenta a importância de cada variável de entrada na rede MLP. É possível observar que as variáveis mais importantes para avaliar a reputação do autor foram "VotosPositivos", "TotalVotos" e "VotosNegativos". Tal resultado se mostra interessante porque a variável "VotosPositivos" é definida por outros usuários da rede, confirmando a boa reputação do autor daquele comentário que recebeu o voto positivo. Por outro lado, a variável menos importante foi "DataCadastro" que indica o tempo, em dias, do cadastro do usuário no site.

Com relação à acurácia de inferência da RNA MLP dentro do conjunto de treinamento e teste, atingiu-se um valor de 62,08% no processo de classificação para os valores numéricos de 0 a 10. No entanto, considerando faixas de valores: 0-3 para baixo, 4-7 para médio e 8-10 para alto, que são normalmente usados em sistemas de reputação, a acurácia da rede atingiu 91,01%.

Tabela 5. Importância das variáveis de entrada na RNA MLP

Variável de entrada	Importância	Importância Normalizada
DataReview	0,106	42,10%
DataCadastro	0,086	34,40%
VotosPositivos	0,252	100,0%
VotosNegativos	0,205	81,70%
TotalVotos	0,245	97,60%
TotalReviewsAutor	0,105	41,80%

5.2. Experimento 2: RNA RBF

Visando apresentar uma alternativa para a arquitetura MLP, executou-se o segundo experimento com uma RNA com funções de bases radial. O melhor ajuste se deu com 11 neurônios na camada escondida e função de ativação *Softmax*. Na camada de saída utilizou-se também o atributo de supervisão "ReputacaoManual" como variável dependente para testar a rede, classificando as 11 notas possíveis dos autores (intervalo de 0 a 10) e função de ativação *Identidade*. Assim como na rede MLP, ressalta-se que os ajustes da rede RBF foram realizados usando a ferramenta SPSS. A topologia com a disposição dos neurônios é semelhante à topologia da rede MLP, com diferença apenas da quantidade de neurônios na camada escondida e das funções de ativação usadas nas camadas escondidas e de saída. O processo de treinamento e teste utilizou o mesmo *Subcorpus I* e o mesmo método de validação *10-fold cross validation*.

A Tabela 6 apresenta a importância de cada variável de entrada usando a rede RBF. Observa-se que os resultados são similares aos apresentados no experimento com a RNA MLP, porém existe uma mudança na ordem das variáveis mais importantes, sendo "TotalVotos", "VotosPositivos" e "VotosNegativos" as mais importantes. Por outro lado, a variável menos importante foi "DataCadastro", assim como na rede MLP.

Tabela 6. Importância das variáveis de entrada na RNA RBF

Variável de entrada	Importância	Importância Normalizada
DataReview	0,127	60,30%
DataCadastro	0,111	52,80%
VotosPositivos	0,210	99,60%
VotosNegativos	0,210	99,50%
TotalVotos	0,211	100,0%
TotalReviewsAutor	0,132	62,60%

Com relação à acurácia de inferência da rede dentro do conjunto de treinamento e teste, atingiu-se um valor de 52,25% no processo de classificação para os valores numéricos de 0 a 10 e 87,36% para as três faixas de valores baixo, médio e alto. Nota-se uma ligeira vantagem da rede MLP sobre a RBF. Dessa forma, decidiu-se usar apenas a primeira arquitetura nos experimentos de comparação entre a abordagem TOP(X) original e a nova abordagem que utiliza uma RNA na dimensão da reputação do autor.

5.3. Experimento 3: Comparação entre as Abordagens

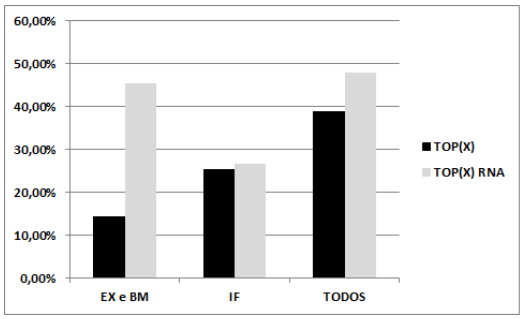
Neste experimento utilizou-se o *Subcorpus II*, contendo 271 comentários anotados com relação a importância, sendo: 17 excelentes (EX), 24 bons (BM), 145 suficientes (SF) e 85 insuficientes (IF). O processo de anotação seguiu recomendações discutidas em [De Sousa et al., 2015].

Para avaliar as abordagens, calculou-se as medidas de Precisão, Cobertura e a medida harmônica Medida-F para cada classe. Destaca-se que essas medidas são normalmente usadas em processos de avaliações na área de aprendizagem de máquina [Powers 2011].

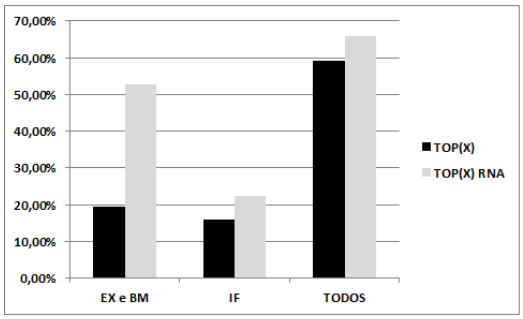
A Figura 5(a) apresenta a comparação baseada na Medida-F entre as duas abordagens, relacionando os comentários positivos em termos de sua importância. É possível observar que abordagem TOP(X) com RNA superou com boa margem a abordagem TOP(X) original nos comentários excelentes e bons. Estes comentários são relevantes pois, normalmente, o usuário procura os melhores comentários para ler e decidir sobre a compra de um produto ou serviço. Desta forma, o usuário poderá focar em um pequeno conjunto de comentários selecionados pela abordagem, gerando um ganho de tempo e esforço na sua avaliação. Ainda na Figura 5(a) observou-se que considerando os comentários ruins (IF) a melhoria foi pouco significativa e, quando se considerou todos os comentários a melhoria da abordagem TOP(X) com RNA foi de cerca de 10%. É importante destacar que os comentários suficientes (SF) não foram apresentados, pois, normalmente, os usuários têm interesse nos melhores comentários ou nos piores (comentários mais críticos).

Destaca-se ainda que os resultados baixos (Medida-F < 50%) podem ser justificados pela dificuldade do especialista em avaliar precisamente, seja na avaliação da reputação do autor (***Subcorpus I***), em que o especialista deve evitar o enviesamento da amostra para não comprometer a capacidade preditiva da RNA, seja na avaliação da importância dos comentários (***Subcorpus II***), que pode gerar inconsistências no processo de avaliação final das abordagens.

A Figura 5(b) apresenta o gráfico comparativo com relação aos comentários negativos e sua respectiva importância, também baseada na Medida-F. A abordagem TOP(X) com RNA MLP mostrou-se superior em desempenho em todos os casos, especialmente na identificação dos comentários excelentes e bons. Já na identificação dos comentários insuficientes e todos agrupados, o desempenho foi ligeiramente superior.



(a) comentários positivos



(b) comentários negativos

Figura 5. Comparação entre as abordagens

5.4. Experimento 4: Outros Algoritmos de AM

Com intenção de observar o comportamento dos dados no domínio de comentários sobre *smartphones*, decidiu-se analisar outros algoritmos de aprendizagem de máquina para o problema de regressão, a saber: *Regressão Linear*, *SVM (versão SMOREG) [Shevade et al., 1999]*, *Tabela de Decisão [Kohavi, 1995]*, *Árvore de Decisão (versão Decision Stump)* e *Processo Guassiano [Mackay, 1998]*. Versões desses algoritmos estão implementadas na ferramenta WEKA[6] e para mais informações a referência George-Nektarios (2013) pode ser consultada. Esses algoritmos foram escolhidos porque eles consideram o problema de regressão linear para a tarefa de predição e utilizam os diversos paradigmas no tratamento do problema: baseado em funções, regras e árvores. Destaca-se que para cada algoritmo, inferiu-se a nota do autor (no intervalo de 0 a 10), a partir das seis variáveis de entrada. De forma semelhante ao que foi feito com as RNAs, em uma segunda análise, observou-se as notas considerando as faixas de valores: 0-3 para baixo, 4-7 para médio e 8-10 para alto.

É importante destacar que este experimento foi realizado com o apoio da ferramenta WEKA, usando a opção de teste *10-fold cross validation*. A Tabela 7 mostra os resultados obtidos com os cinco algoritmos, bem como os resultados das redes neurais MLP e RBF. A métrica de avaliação observada foi a acurácia de inferência dos modelos. A coluna *Nota 0-10* significa que a nota do autor foi inferida para os valores variando de 0 a 10, enquanto que a coluna *Faixas* significa que o valor da nota foi considerando as faixas: **baixo** (0-3), **médio** (4-7) e **alto** (8-10).

Tabela 7. Acurácia dos algoritmos de AM analisados

Modelos/Algoritmos	***10-Fold Cross Validation***	
	Nota 0-10	***Faixas***
Regressão Linear	41,29%	84,27%
SMOReg	32,02%	84,55%
Tabela de Decisão	43,82%	83,71%
Decision Stump	42,98%	77,81%
Processo Gaussiano	41,01%	85,11%
RNA MLP	62,08%	91,01%
RNA RBF	52,25%	87,26%

Observa-se que todos os cinco algoritmos de regressão utilizados possuem acurácia inferior os modelos de redes neurais, considerando tanto os valores de 0 a 10, quanto as faixas de valores. Portanto, os dados mostram que as redes neurais são melhores aproximadores da reputação do autor no problema de avaliação de comentários Web sobre produtos.

6. Análise de Erros

Esta seção discute os motivos dos resultados obtidos nos experimentos terem sido pouco expressivos. Inicialmente, um dos fatores que pode ter contribuído para os baixos

[6] https://www.cs.waikato.ac.nz/ml/weka/

resultados foi o processo de anotação manual do *Corpus*, que foi realizado de três formas diferentes: quanto à polaridade, quanto à reputação do autor e quanto à importância.

Em processos de avaliação realizados por seres humanos, os especialistas podem divergir quanto às suas decisões, devido à natureza subjetiva das opiniões coletadas. Neste sentido, algumas pesquisas demonstram que em anotações feitas por humanos, dificilmente o consenso é maior que 75% [Bruce and Wiebe 1999, Ku et al. 2006, Pang et al. 2002, Wiebe et al. 2005].

A Figura 6 apresenta um diagrama com o quantitativo das classes de importâncias anotadas manualmente e das classes inferidas pelas abordagens TOP(X) Original e TOP(X) com RNA. Pode-se observar que, em muitos casos, as duas abordagens conseguem atingir um bom número de acertos na classificação, no entanto, em outros existem muitos erros na inferência. Por exemplo, as duas abordagens apresentam bom desempenho para inferir os comentários "suficientes", porém para os comentários "insuficientes", o desempenho de ambas é ruim. Outro aspecto a ser observado é que as abordagens falham ao inferir uma classe corretamente por uma pequena margem, por exemplo, o Sistema *Fuzzy* da Abordagem TOP(X) infere uma saída numérica 6,9 classificando o comentário como "bom", porém a anotação manual o classificou como "excelente", gerando um erro. Mais informações sobre as configurações do Sistema *Fuzzy* utilizado no abordagem TOP(X), incluindo funções de pertinências das variáveis linguísticas e a base de regras, podem ser consultadas em [De Sousa et al. 2015].

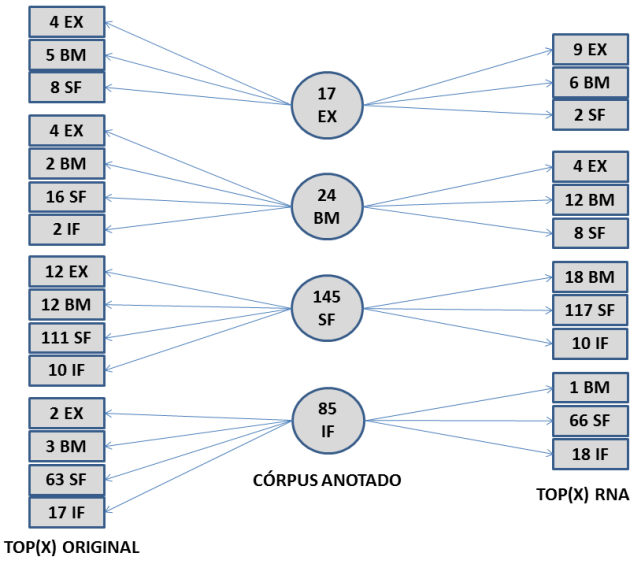


Figura 6. Análise de erros

Além dos problemas citados, suspeita-se que outros aspectos podem causar inconsistências no processo de classificação das abordagens. Por exemplo, imprecisões das ferramentas que implementam as técnicas de AM, enviesamento no processo de treinamento/teste da Rede Neural, inconsistências na elaboração das regras de pertinência e variáveis linguísticas do Sistema *Fuzzy*.

Por fim, os erros mencionados podem ser amenizados com a sistematização do processo de anotação do *Corpus* e a realização de novos experimentos.

7. Conclusões e Trabalhos Futuros

Sistemas de recomendação baseados em dados abertos são de suma importância para a sociedade e para as organizações em geral e representam um dos desafios da área de Sistemas de Informação (SI). Nosso estudo pode ser utilizado diretamente para a criação de soluções de problemas do mundo real que envolvam a participação dos cidadãos. À luz de nosso conhecimento, não existe nenhuma solução comercial para criar um *ranking* de comentários sobre produtos e serviços, analisando as características textuais dos relatos e as informações dos autores do comentário.

Neste artigo foi apresentado um estudo sobre métricas para definir a reputação do autor de comentários em sites de comparação de preços de produtos. De forma geral, o modelo proposto representa uma adaptação da abordagem TOP(X) [De Sousa et al. 2015], com ênfase na dimensão reputação do autor. O estudo foi conduzido através da aplicação de redes neurais para inferir a reputação dos autores dos comentários e descobrir a importância de cada medida de entrada.

Realizou-se dois experimentos com as RNAs MLP e RBF, usando o ***Subcorpus I*** para realizar o treinamento da rede. Os resultados obtidos apresentaram similaridades entre as redes quanto à indicação da importância das variáveis de entrada. Observou-se que a quantidade de votos positivos que um autor recebe tem um peso significativo em sua reputação, sendo considerada a principal medida para avaliar a reputação do autor no contexto analisado. Com relação ao desempenho dos modelos, a rede RBF atingiu 87,36% de acurácia, enquanto que a rede MLP atingiu 91,01%. Assim, a rede MLP foi utilizada nos demais experimentos de nossas pesquisas.

Em um terceiro experimento comparou-se as abordagens TOP(X) original e TOP(X) com RNA MLP, utilizando o ***Subcorpus II*** e observando a Medida-F (média harmônica entre a precisão e a cobertura). Com foco nos comentários excelente e bons, a nova abordagem apresentou resultados significativamente superiores. Conclui-se, então, que tal abordagem pode auxiliar os usuários na busca por produtos ou serviços, reduzindo o tempo e esforço gastos no processo. Adicionalmente, um quarto experimento foi realizado para analisar os dados, usando outros algoritmos de AM. Os cinco algoritmos avaliados obtiveram acurácia inferior aos modelos das RNAs.

Como trabalhos futuros, pretende-se: i) aplicar a nova abordagem em um *Corpus* maior, realizando um processo mais extenso de anotação manual; e ii) investigar o impacto de reputação do autor em notícias falsas (*fake news*), pois sabe-se que existem alguns artifícios utilizados para potencializar o alcance de uma notícia ou comentário, bem como impulsionar a reputação de um autor em mídias sociais.

Referências

Adler, B. T. and De Alfaro, L. (2007). "A content-driven reputation system for the Wikipedia". In *Proc. of the Int. Conference on World Wide Web*, pages 261–270. ACM. [GS Search]

Adler, B. T., De Alfaro, L., and Pye, I. (2010). "Detecting wikipedia vandalism using wikitrust". Notebooks papers of CLEF. [GS Search]

Aggarwal, A. and Kumaraguru, P. (2014). “Followers or phantoms? an anatomy of purchased twitter followers”. *CoRR*. [GS Search]

Bruce, R. F. and Wiebe, J. M. (1999). “Recognizing subjectivity: A case study of manual tagging”. *Natural Language Engineering*, v. 5, p. 187–205. [GS Search]

Cappelletti, R. and Sastry, N. (2012). “IARank: Ranking users on twitter in near real-time, based on their information amplification potential”. In *SocialInformatics*, pages 70–77. [GS Search]

Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). “Clues for detecting irony in user-generated contents: Oh...!! it’s ”so easy”;-)”. In *Proc. of the Int. Workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56. [GS Search]

De Sousa, R. F., Rabelo, R. A. L., and Moura, R. S. (2015). “A fuzzy system-based approach to estimate the importance of online customer reviews”. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. [GS Search]

George-Nektarios, T. (2013). "Weka Classifiers Summary", [online] Available: www.academia.eduI5167325IWeka_Classifiers_Summary. [GS Search]

Gonçalves, P., Dalip, D., Reis, J., Messias, J., Ribeiro, F., Melo, P., Araújo, L., Gonçalves, M., and Benevenuto, F. (2015). “Bazinga! caracterizando e detectando sarcasmo e ironia no Twitter”. In *Proc. of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [GS Search]

Halfaker, A., Kittur, A., Kraut, R. and Riedl, J. (2009). “A jury of your peers: Quality, experience and ownership in Wikipedia”. In *Proc. of the Int. Symposium on Wikis and Open Collaboration*, pages 15:1–15:10. ACM. [GS Search]

Hamilton, R., Vohs, K. D. and McGill, A. L. (2014). “We’ll be honest, this won’t be the best article you’ll ever read: The use of dispreferred markers in word-of-mouth communication”. *Journal of Consumer Research*, 41(1):197 – 212. [GS Search]

Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M., Pardo, T. and Aluísio, S. (2014). “A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words”. In *Proc. of the Int. Conference on Language Resources and Evaluation (LREC’14)*. [GS Search]

Hoisl B., Aigner W., Miksch S. (2007) “Social Rewarding in Wiki Systems – Motivating the Community”. In: Int. Conf. on Online Communities and Social Computing. Springer. [GS Search]

Jackson, P. and Moulinier, I. (2007). “*Natural language processing for online applications: Text retrieval, extraction and categorization”*. John Benjamins, Amsterdam. [GS Search]

Jones, C., Hesterly, W. S. and Borgatti, S. P. (1997). “A general theory of network governance: Exchange conditions and social mechanisms”. *The Academy of Management Review*, Academy of Management, v. 22, n. 4, p. 911–945. [GS Search]

Kohavi, R. (1995). “The Power of Decision Tables”. In: 8th European Conference on Machine Learning,174-189, 1995. [GS Search]

Ku, L.-W., Liang, Y.-T. and Chen, H.-H. (2006) "Opinion extraction, summarization and tracking in news and blog Corpora". In: *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*. [GS Search]

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). "What is Twitter, a social network or a news media?" In *Proc. of the Int. Conf. on World Wide Web*. ACM. [GS Search]

Li, B., Li, RH, King, I., Lyu MR and Yu, JX (2014). "A topic-biased user reputation model in rating systems". *Knowledge and Information Systems*. [GS Search]

Li, M., Huang, L., Tan, C. and Wei, K. (2013). "Helpfulness of online product reviews as seen by consumers: Source and content features". *Int. J. Electronic Commerce*, 17(4):101–136. [GS Search]

Liu, B. (2011). "*Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*". Springer-Verlag New York, Inc., Secaucus, NJ, USA. [GS Search]

Mackay, D. (1998). "Introduction to Gaussian Processes". Dept. of Physics, Cambridge University, UK. [GS Search]

Pang, B., Lee, L. and Vaithyanathan, S. (2002) "Thumbs up?: Sentiment classification using machine learning techniques". In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. [GS Search]

Powers, D. M. W. (2011). "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation". *Journal of Machine Learning Technologies*, 2(1):37–63. [GS Search]

Shevade, S., Keerthi, S., Bhattacharyya, C., Murthy, K. (1999). "Improvements to the SMO Algorithm for SVM Regression". In: IEEE Transactions on Neural Networks. [GS Search]

SPSS (2007). *SPSS for Windows, Version 16.0*. Chicago USA: [s.n.].

Wang, A. H. (2010). "Don't follow me: Spam detection in Twitter". In *Proc. of the Int. Conference on Security and Cryptography (SECRYPT)*, pages 1–10. [GS Search]

Weitzel, L., de Oliveira, J. P. and Quaresma, P. (2014). "Measuring the reputation in user-generated-content systems based on health information". *Procedia Computer Science*, 29:364 – 378. [GS Search]

Wiebe, J., Wilson, T. and Cardie, C. (2005) "Annotating expressions of opinions and emotions in language". In: Language Resources and Evaluation, 39 2-3. [GS Search]

Weng, J., Lim, E.-P., Jiang, J. and He, Q. (2010). "Twitterrank: Finding topic-sensitive influential Twitterers". In *Proc. of the Int. Conference on Web Search and Data Mining*, pages 261–270. ACM. [GS Search]

Wöhner, T., Köhler, S., and Peters, R. (2011). "Automatic reputation assessment in Wikipedia". In *Proc. of the Int. Conference on Information Systems*. [GS Search]

Zhao, H., Ye, S., Bhattacharyya, P., Rowe, J., Gribble, K. and Wu, S. F. (2010). "Socialwiki: Bring order to wiki systems with social context". In: Int. Conf. on *Social Informatics*. [GS Search]