Aplicando Redes Neurais sem Peso em Mineração de Opinião no Twitter

Fabio Rangel, Fabrício Firmino, Samara Alves, Jonice de Oliveira Sampaio, Priscila Machado Vieira Lima
Programa de Pós-Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro
{fabiorangel, firminodefaria, jonice}@ufrj.br, samara.alvarez88@gmail.com, priscila.lima@nce.ufrj.br

***Resumo*—Redes Sociais *online*, como por exemplo o Twitter, são amplamente utilizadas pelos seus usuários para expressar opiniões sobre questões que os cercam diariamente. Por meio da aplicação de algoritmos de classificação é possível a identificação de padrões capazes de categorizar as mensagens oriundas de mídias sociais sob diversos pontos de vista. Os classificadores WiSARD podem ser utilizados para classificação de documentos textuais no modelo *bag-of-words*. O presente trabalho visa comparar a aplicação do classificador WiSARD com outros classificadores amplamente utilizados na mineração de opinião no contexto de documentos textuais oriundos de mídias sociais *online*. O classificador WiSARD apresentou acurácia próxima a do SVM e superior a do *Naive Bayes* nos dois *datasets* de teste utilizados.**

***Palavras-chave*—Opinion Mining, WiSARD, Weightless Neural Networks, Support Vector Machines, Naive Bayes, Bag of Words.**

I. Introdução

Redes sociais *online* tornaram-se uma importante ferramenta de comunicação para a sociedade. Por meio dessas mídias, usuários compartilham suas opiniões sobre diversos assuntos. Para ilustrar a relevância do conteúdo publicado nas redes sociais *online*, a revista The New York Times apresentou um artigo destacando a importância dessas redes durante as eleicões presidenciais nos Estados Unidos em 2012 [1]. Outra importante utilização das mídias sociais está relacionada ao *marketing* digital, onde redes sociais *online* aproximaram empresas e consumidores. O *company branding* e o *word-of-mouth advertising* foram destacados pela Forbes como principais benefícios das mídias sociais em uma pesquisa de tendências para o ano de 2014 [2].

"O que as pessoas pensam" é uma importante informação, para a maioria das pessoas, durante um processo de tomada de decisão [3], e a mineração de opinião tem como objetivo responder essa questão apoiada em técnicas de aprendizado de máquina. Tratando-se de documentos textuais oriundos de redes sociais *online*, tais técnicas tem como objetivo encontrar padrões nesses documentos e categorizá-los automaticamente sob um determinado ponto de vista. Existem diversas abordagens para extrair padrões de dados. Dentre as abordagens supervisionadas mais utilizadas no cenário de mídias sociais *online*, estão o *Support Vector Machine* (SVM) e o *Naive Bayes*.

Devido à disponibilidade do seu conteúdo, o Twitter é uma das redes sociais *online* mais estudadas. Muitos desses estudos visam identificar padrões de mensagens compartilhadas na rede com objetivo de classificá-las automaticamente. Por exemplo, o Twitter foi utilizado como fonte de dados para analisar a opinião pública a respeito das manifestações ocorridas em junho de 2013 no Brasil [4]. Os *tweets* foram classificados automaticamente como apoiando ou repudiando as manifestações, e o classificador utilizado no trabalho foi o *Naive Bayes*. Outro trabalho apresenta como objetivo identificar o alinhamento político dos usuários do Twitter por meio da classificação das mensagens desses usuários. Para isso, foi utilizado o SVM para categorizar as mensagens em direita ou esquerda [5].

Embora o *Naive Bayes* e o SVM sejam classificadores amplamente utilizados no contexto do Twitter, essas abordagens não possuem uma solução prática para a inserção de novos documentos de treino a partir de um modelo já treinado, sendo necessário treinar o modelo novamente com todos os exemplos. Seria conveniente um modelo que pudesse treinar um novo exemplo aproveitando o treino anteriormente existente. Um modelo que se adapta a essa característica é a Rede Neural sem Peso WiSARD. Essa adaptação é interessante também para o próprio cenário de redes sociais *online*, onde o contexto muda constantemente e as palavras associadas a uma determinada polaridade podem ser desassociadas. Além disso, o modelo de representação de documentos textuais através de vetores de *features* (*bag-of-words*) é próprio para ser aplicado no classificador WiSARD, sendo necessário apenas transformar os valores das *features* em *bits* (binarização). Dessa forma, o padrão de entrada pode ser mapeado para o modelo *bag-of-words*.

O presente trabalho apresenta uma comparação entre o classificador WiSARD e os classificadores *Naive Bayes* e SVM, no cenário de documentos textuais oriundos de mídias sociais *online*. Para comparar esses classificadores, foram utilizados medidas de acurácia e desvio padrão do resultado de 30 aplicações do 10-Fold Cross Validation em dois diferentes *datasets*: OMD e STS-Gold. Dentre os resultados encontrados, a WiSARD se comportou de forma satisfatória, empatando com o SVM em um *dataset* e se aproximando da acurácia do SVM em outro. Em todos os resultados de acurácia a WiSARD apresentou valores melhores que o *Naive Bayes*.

O restante do trabalho se subdivide em: Conceitos Básicos, onde é apresentado o classificador WiSARD e os classificadores de comparação; Metodologia, onde são descritos os *datasets* de teste, o pré-processamento e o experimento;

Experimentos e Análises, onde os resultados são apresentados e debatidos; Conclusão e Trabalhos Futuros.

II. Conceitos Básicos

A. Classificador WiSARD

O WiSARD (Wilkie, Stonham & Aleksander's Recognition Device) é um tipo de Rede Neural sem Peso [6]. O formato de entrada é uma cadeia de *bits*, isso porque o reconhecimento das imagens originalmente era feito em preto-e-branco. Assim sendo, objetos que não estejam nesse padrão devem ser convertidos (binarizados) [7].

O conteúdo da memória de acesso aleatório (RAM) é modificado com o objetivo de acumular aprendizado. Contudo, um problema do modelo é a perda da capacidade de generalização, pois ela reconhece apenas padrões já apresentados na fase de treinamento. Para tentar contornar esse problema, os neurônios da memória são organizados no discriminador. Cada um desses discriminadores é responsável por reconhecer exemplos de uma determinada classe, ou seja, cada um deles tem capacidade de reconhecer uma classe de padrões que já foi treinada [8].

1) Treinamento: O treinamento da WiSARD, representado na Figura 1, é supervisionado, portanto se faz necessário informar a que classe um padrão pertence. Sendo assim, nessa etapa se define o discriminador que estará associado a cada classe. De início, o conteúdo das posições das RAMs é inicializado com zero e durante o treinamento eles são atualizados.

É importante ressaltar que as escolhas dos n *bits* nos padrões de entrada é feita de forma aleatória, e que essa escolha deve ser sempre a mesma para todos os discriminadores e padrões de entrada, visando assim garantir a integridade do treinamento na etapa de reconhecimento. Dessa forma, é incrementado em uma unidade o conteúdo que tiver como endereçamento esses n bits [8].

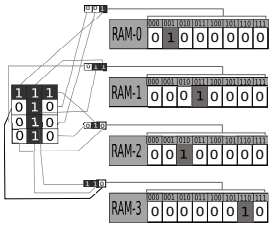

Figura 1. Treino da WiSARD com um exemplo da letra T [9].

2) Reconhecimento: Na etapa do reconhecimento, apresentado na Figura 2, a rede analisa a resposta de todos os discriminadores e escolhe o que ativou o maior número de neurônios [8]. Esses neurônios produzem saídas binárias: 1 se a posição que foi indicada pelo endereço que a RAM recebe estava marcado; e 0, caso contrário [10]. O placar de um discriminador é a soma das respostas de todas as RAMs desse discriminador [8]. A Figura 3 apresenta os placares de cada discriminador na etapa de reconhecimento.

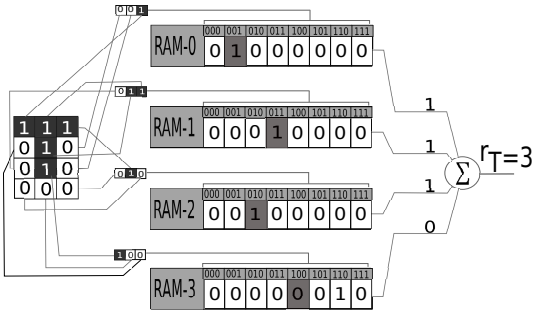

Figura 2. Modelo do Classificador WiSARD [9].

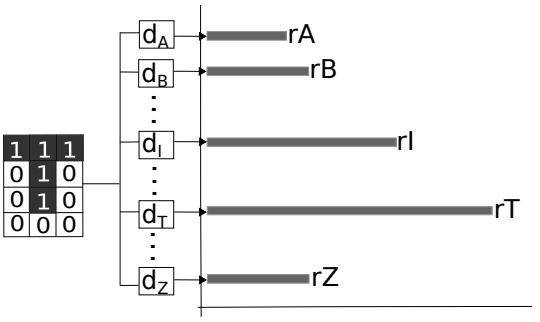

Figura 3. Classificação da WiSARD com placar dos Classificadores [9].

B. Mecanismo de Bleaching

Um problema encontrado na WiSARD é a saturação da rede: perda da capacidade de reconhecimento quando os discriminadores recebem treinamento em excesso. Essa saturação implica em empates que acontecem nos placares dos discriminadores. Uma maneira encontrada para contornar esse problema foi o mecanismo de *Bleaching*, representado na Figura 4. A fase de treinamento com esse mecanismo é alterada no seguinte aspecto: ao invés de armazenar 1, ele incrementa $+1$ no conteúdo da memória responsável, sendo assim, terminada a fase de treinamento os valores do conteúdo da memória irão variar de 0 a X, onde X é o número de padrões de treinamento [10]. Esse mecanismo permite obter maiores detalhes sobre a etapa de treinamento e desta forma saber quais subconjuntos foram acessados dentro do discriminador.

Existem duas formas de *Bleaching*: o homogêneo e o heterogêneo. O primeiro é quando o mesmo valor de *Bleaching* é usado para todos os discriminadores. Já o segundo varia de acordo com a quantidade de aprendizado de cada discriminador. A análise de qual mecanismo usar requer testes, haja vista que ela não é automática [8]. A Figura 3 ilustra a classificação dos discriminadores com o emprego do mecanismo de *Bleaching*.

C. Classificadores de Comparação

Para avaliar a acurácia da WiSARD quando aplicada na classificação automática são necessários outros algoritmos de classificação. Para tanto, decidiu-se utilizar os algoritmos *Naive Bayes* e SVM que utilizam o modelo *bag-of-words* em sua classifiçacão. De uma forma geral, o modelo *bag-of-words* é uma representação muito simples, onde é ignorada a ordem

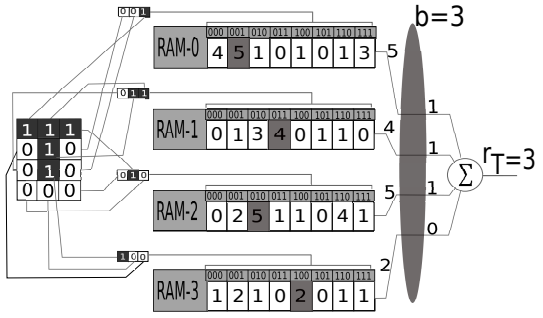

Figura 4. WiSARD com *Bleaching* [9]

das palavras nos documentos, limitando-se a contar o número de ocorrências de cada palavra (*feature*) [11]. Outro fator relevante quanto à escolha dos algoritmos é o fato de serem amplamente utilizados em diversos trabalhos no mesmo tema [12].

Naive Bayes é um tipo de classificador Bayesiano. Esse algoritmo assume que todos os atributos dos exemplos são independentes entre si dado o contexto de uma classe [13]. Apesar de sua simplicidade e o fato de sua suposição de independência condicional claramente não se sustentar em situações no mundo real, a categorização textual do *Naive Bayes* tende a ser realizada surpreendentemente bem [14].

Support Vector Machine (SVM) [15] possui uma forte fundamentação teórica e expressivo sucesso empírico. Este é aplicado, por exemplo, em tarefas como reconhecimento de dígitos escritos a mão, reconhecimento de objetos e classificação textual [16]. Atualmente há inúmeros trabalhos que ainda utilizam o SVM para classificar diferentes tipos de dados. Define-se o mesmo como um classificador não linear, inicialmente formulado por Vapnik [17], operando em um espaço de dimensão maior do que o espaço original do conjunto de *features* da base de dados [18].

III. Metodologia

A. Datasets de Teste

Esta seção apresenta informações sobre os *datasets* utilizados para medir a acurácia dos classificadores. Importante perceber que as duas bases utilizadas possuem número de classes diferentes, sendo assim, haverá uma diferença na acurácia entre os mesmos classificadores nas duas bases. A Tabela I apresenta o número de documentos por classe em cada uma das bases.

Tabela I
Número de Documentos por Classe nos Datasets.

Data Set	Positive	Negative	Mixed	Other
OMD	844	1582	289	523
STS-Gold	632	1402	-	-

1) Obama-McCain Debate (OMD): O *dataset* Obama-McCain Debate (OMD) foi construído de 3.238 tweets durante o primeiro debate presidencial dos EUA na TV, em setembro de 2008 [19]. Para a classificação dos sentimentos desses tweets foi utilizado o Amazon Mechanical Turk, no qual cada tweet é classificado por um ou mais examinadores como as seguintes classificações: positivo; negativo; misto; e outros. A categoria *outros* são àqueles que não puderam ser classificados. O OMD é comumente utilizado como base de teste em diferentes trabalhos [20]. Por exemplo, esse dataset foi utilizado em um trabalho onde foram reunidas várias abordagens de classificação via *label propagation* além de realizar um estudo para verificar se a estrutura do grafo de relacionamentos do Twitter poderia melhorar a classificação de polaridade nos documentos [21].

2) STS-Gold: STS-Gold é um *dataset* cuja criação foi baseada em outro *dataset* anteriormente existente chamado STS (Stanford Twitter Sentiment). A base STS consiste em 1,6 milhões de *tweets* anotados como positivos ou negativos baseados em *emoticons* [20]. O STS-Gold é um recorte do STS onde manualmente foram anotados os sentimentos dos *tweets* e as entidades presentes no mesmo. Embora o *dataset* STS-Gold seja recente (2013), é possível encontrar trabalhos na literatura utilizando essa base. Por exemplo, o STS-Gold foi utilizado em um trabalho onde é apresentado um método de *feature selection* baseado na descoberta automática de *stop words* visando melhorar a acurácia da classificação automática [22].

B. Pré-Processamento

O pré-processamento aplicado no presente trabalho consistiu nas seguintes etapas: (i) transformar todas as letras para minúsculo; (ii) remover as pontuações; (iii) remover links; (iv) remover *mentions* (específico do Twitter); (v) *stemming*. Este último é um método amplamente utilizado de padronização de palavras projetadas para permitir o *text matching* de termos morfológicos relacionados, com objetivo de remover os afixos e, dessa forma, reduzir a palavra para o seu radical [23]. O algoritmo de *stemming* aplicado foi o Porter Stemmer [24] que remove os sufixos dos termos em inglês.

C. Acurácia e Validação Cruzada

Para medir a acurácia dos algoritmos, foi adotado o *K-Fold Cross Validation* [25]. Quando essa métrica de avaliação é utilizada, os dados são divididos em k segmentos igualmente (ou quase igualmente) distribuídos. Subsequentemente, k iterações de treinamento e validação são realizadas de tal modo que, em cada iteração apenas um segmento diferente dos dados é utilizado na validação, enquanto os $k-1$ segmentos restantes são utilizados no treinamento [25]. No presente trabalho, a acurácia foi medida calculando a média dos resultados de 30 diferentes processos de *K-Fold Cross Validation*n com k igual a 10. Além disso, calculamos da mesma forma o desvio padrão.

IV. Experimentos e Análises

A. Verificando Número de Features

Para verificar a dimensão utilizada na classificação, calculou-se o número de *features* em cada *dataset*. É possível encontrar na literatura que a dificuldade da classificação em dimensões elevadas é causada pela existência de muitos *noise*

features, que não contribuem para a redução da taxa de *misclassification* [26]. Embora essa afirmação possa ser uma verdade, no presente trabalho não foram aplicadas técnicas de redução de dimensionalidade, tais como *Document Frequency Thresholding* [27] ou *Singular Value Decomposition* [28]. A Tabela II apresenta o número de *features* em cada *dataset* após o pré-processamento. É possível perceber que ambos *datasets* possuem dimensões elevadas, passando de 3000 *features*.

Tabela II
NÚMERO DE FEATURES POR DATASET APÓS O PRÉ-PROCESSAMENTO

Dataset	Número de Documentos	Número de Features
OMD	3238	5436
STS-Gold	2034	5411

B. Dataset OMD

Calculou-se a acurácia média e o desvio padrão de cada classificador no *dataset* OMD. Nos primeiros experimentos, verificou-se que a WiSARD com *Bleaching* apresentava-se menos acurada que a WiSARD sem *Bleaching*. Como o *Bleaching* é uma técnica para melhorar a acurácia da WiSARD, esse problema chamou a atenção. Foi verificado que, devido à esparsidade do modelo de representação *bag-of-words*, as posições de memória da WiSARD cujos *bits* estavam todos zerados apresentavam os maiores valores após a fase de treinamento. O *Bleaching* acabava por ressaltar essas posições e apagar as posições cujos valores refletiam a presença de uma determinada *feature*. Para solucionar esse problema, foram retiradas de todas as memórias as posições cujos *bits* eram todos zerados. Essa mudança não somente melhorou a a acurácia da WiSARD com *Bleaching* como melhorou também a acurácia da WiSARD sem *Bleaching*.

A Figura 5 apresenta a acurácia média dos classificadores utilizados comparando com os classificadores WiSARD. Os classificadores WiSARD foram utilizados com 8, 12 e 16 *bits*, apresentando valores de acurácia média parecidos. A WiSARD com *Bleaching* se comportou com acurácia próxima ao do SVM, e o *Naive Bayes* não apresentou valores competitivos para essa base. A WiSARD sem *Bleaching* apresentou valores maiores que o *Naive Bayes* para todos os *bits* utilizados. Embora os valores da acurácia média estejam próximos de 50%, isso se deve ao fato deste *dataset* possuir 4 classes. A classificação textual com mais do que 2 classes ainda apresenta dificuldades mesmo para abordagens consolidadas como o SVM [29].

Além do cálculo da acurácia média, o desvio padrão foi calculado. O desvio padrão da acurácia é importante para verificar o quão preciso é aquele classificador. A Figura 6 apresenta o desvio padrão para cada um dos classificadores no *dataset* OMD. É possível perceber que não existe tanta diferença entre os classificadores, todos apresentam o desvio entre 2% e 3%.

C. Dataset STS-Gold

Para o *dataset* STS-Gold, utilizou-se a mesma modificação na WiSARD descrita na Seção IV-B. Diferente do *dataset*

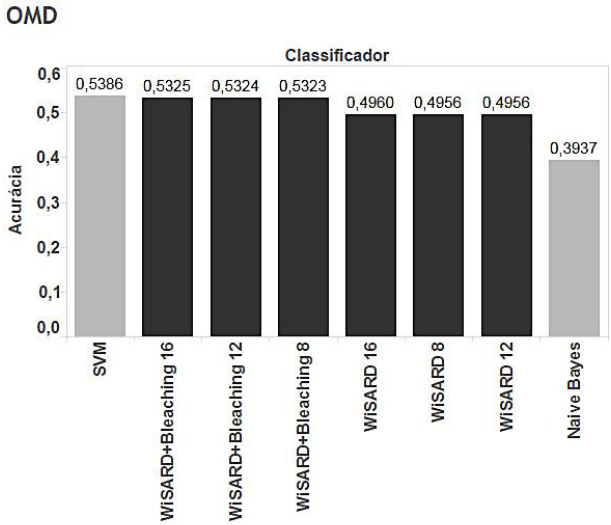

Figura 5. Acurácia Média dos Classificadores no Dataset OMD.

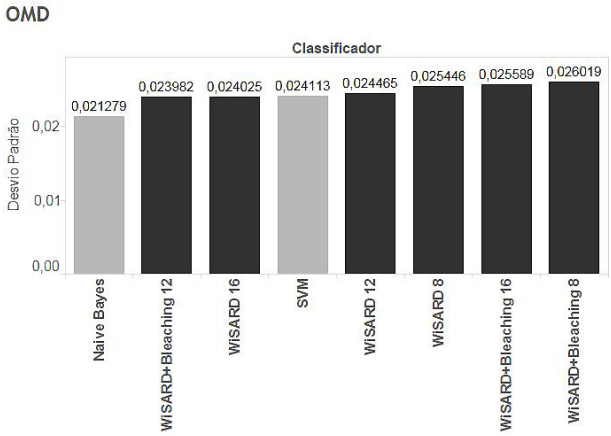

Figura 6. Desvio Padrão da Acurácia entre os Classificadores no Dataset OMD.

OMD, o STS-Gold possui apenas duas classes, permitindo aos classificadores apresentarem maiores valores de acurácia média. A Figura 7 apresenta a acurácia média dos classificadores no *dataset* STS-Gold. Para esse *dataset*, a WiSARD com 8 *bits* apresentou uma acurácia média de 77% enquanto o SVM obteve 81%. Embora essa diferença seja considerável, é possível encontrar na literatura que humanos concordam em cerca de 72% em uma classificação manual, tornando o resultado da WiSARD com 8 *bits* um valor aceitável em uma classificação com duas classes. Pode-se perceber também que a WiSARD, independente do *Bleaching* ou do número de *bits*, se manteve mais acurada que o *Naive Bayes*. Esse é um resultado muito importante pois o *Naive Bayes* é amplamente utilizado na classificação textual no cenário de mídias sociais *online*.

Também calculou-se o desvio padrão da acurácia no *dataset* STS-Gold e o resultado pode ser visto na Figura 8. O SVM se manteve o mais acurado e também apresentou o menor desvio padrão da acurácia. A WiSARD com 8 *bits* não apresentou um desvio padrão distante, ainda abaixo dos 3%.

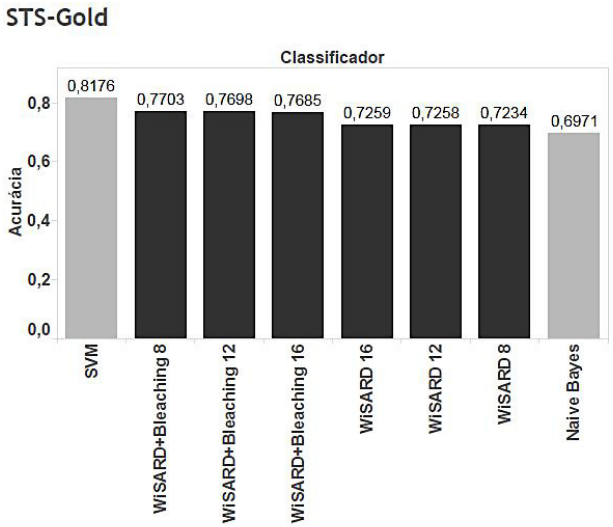

Figura 7. Acurácia Média dos Classificadores no Dataset STS-Gold.

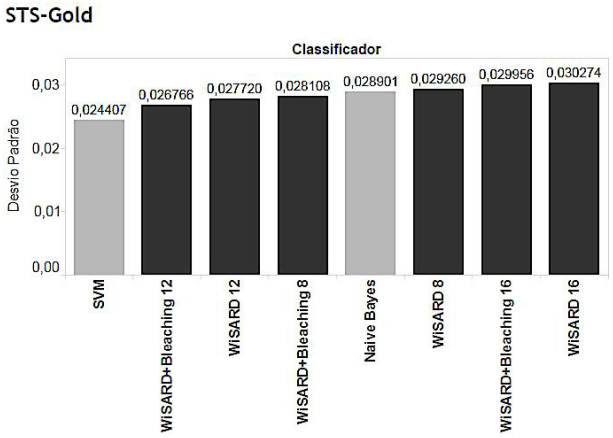

Figura 8. Desvio Padrão da Acurácia entre os Classificadores no Dataset STS-Gold.

V. Conclusões e Trabalhos Futuros

O presente trabalho apresentou uma comparação entre o classificador WiSARD, SVM e o *Naive Bayes* no cenário de classificação textual de documentos oriundos de mídias sociais *online*. A acurácia desses classificadores foi comparada utilizando dois *datasets* que estiveram presentes em outros trabalhos sobre classificação textual no mesmo cenário: OMD e STS-Gold. Juntamente com o classificador WiSARD, foi utilizado o mecanismo de *Bleaching* para contornar o problema de saturação existente no modelo da WiSARD. Pode-se afirmar com base nos resultados obtidos que a WiSARD apresentou uma acurácia satisfatória, chegando a empatar com o SVM no *dataset* OMD. Além disso, no *dataset* STS-Gold, a WiSARD chegou a 77% de acurácia enquanto o SVM apresentou 81%. Embora essa diferença seja considerável, deve-se notar a vantagem que a WiSARD tem sobre outros modelos de classificação onde é possível adicionar novos exemplos de treino sem ter a necessidade de treinar todo conjunto novamente. Além disso, o *Naive Bayes*, que é um classificador amplamente utilizado no contexto de mídias sociais *online* apresentou acurácia menor que a da WiSARD, independente do *Bleaching* ou do número de *bits*. O desvio padrão da acurácia também foi utilizado para comparar os classificadores, porém todos apresentaram resultados entre 2% e 3%, sendo assim, nenhum classificador se destacou nesse parâmetro de comparação.

O presente trabalho também apresentou uma abordagem para adequar a WiSARD ao modelo *bag-of-words* onde foi possível melhorar a acurácia desse modelo de classificador. A abordagem consiste em não considerar as posições de memória onde todos os *bits* estejam zerados. Essa modificação melhora a classificação visto que a representação de um documento no modelo *bag-of-words* é um vetor esparso, e o número de *hits* na posição de memória onde todos os *bits* estão zerados acaba por ser muito maior do que os outros, prejudicando a classificação nesse modelo. Pelo fato da implementação da WiSARD neste trabalho se tratar ainda de um protótipo, e ambos SVM e *Naive Bayes* possuírem implementações maduras e otimizadas, não se propôs a realizar aqui uma comparação de tempo de processamento.

Dentre os trabalhos futuros, pode-se destacar a utilização de técnicas para redução de dimensionalidade com finalidade de comparar a acurácia dos classificadores em dimensões menores. Além disso, pretende-se verificar quais dessas técnicas poderiam ser utilizadas para melhorar a classificação da WiSARD. Dentre as possíveis técnicas, pode-se destacar: *Document Frequency Thresholding* e *Singular Value Decomposition* (para selecionar *features*). Outro possível trabalho futuro é realizar uma comparação do tempo de processamento da WiSARD com outros classificadores em uma abordagem de classificação *on-line*.

Agradecimentos

CNPq, FAPERJ e CAPES.

Referências

[1] The New York Times, "The presidential campaign on social media," 2012. [Online]. Available: http://www.nytimes.com/interactive/2012/10/08/technology/campaign-social-media.html?_r=0

[2] Forbes, "The top 7 social media marketing trends that will dominate 2014," 2013. [Online]. Available: http://www.forbes.com/sites/jaysondemers/2013/09/24/the-top-7-social-media-marketing-trends-that-will-dominate-2014/

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] T. França and J. Oliveira, "Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013," in *III Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*. SBC, 2014.

[5] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 192–199.

[6] I. Aleksander, M. De Gregorio, F. M. G. França, P. M. V. Lima, and H. Morton, "A brief introduction to weightless neural systems." in *ESANN*. Citeseer, 2009.

[7] D. de Oliveira Cardoso, "Uma arquitetura para agrupamento de dados em fluxo contínuo baseada em redes neurais sem pesos," Ph.D. dissertation, Universidade Federal do Rio de Janeiro, 2012.

[8] L. C. Bandeira and F. M. G. França, "Nc-wisard: Uma interpretação sem pesos do modelo neural neocognitron," Master's thesis, Universidade Federal do Rio de Janeiro (UFRJ), 2010.

[9] H. C. Carneiro, F. M. França, and P. M. Lima, "Multilingual part-of-speech tagging with weightless neural networks," *Neural Networks*, vol. 66, no. 0, pp. 11 – 21, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608015000465

[10] B. P. Grieco, P. M. Lima, M. De Gregorio, and F. M. França, "Producing pattern examples from "mental" images," *Neurocomputing*, vol. 73, no. 7, pp. 1057–1064, 2010.

[11] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, 2006.

[12] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093 – 1113, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2090447914000550

[13] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[14] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine learning: ECML-98*. Springer, 1998, pp. 4–15.

[15] V. N. Vapnik and S. Kotz, *Estimation of dependences based on empirical data*. Springer-Verlag New York, 1982, vol. 41.

[16] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1162/153244302760185243

[17] V. Vapnik, "The nature of statical learning theory," 1995.

[18] D. N. Sotiropoulos, C. D. Kounavis, P. Kourouthanassis, and G. M. Giaglis, "What drives social sentiment? an entropic measure-based clustering approach towards identifying factors that influence social sentiment polarity," in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*. IEEE, 2014, pp. 361–373.

[19] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: Understanding community annotation of uncollected sources," in *Proceedings of the First SIGMM Workshop on Social Media*, ser. WSM '09. New York, NY, USA: ACM, 2009, pp. 3–10. [Online]. Available: http://doi.acm.org/10.1145/1631144.1631148

[20] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold," 2013.

[21] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proceedings of the First Workshop on Unsupervised Learning in NLP*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 53–63. [Online]. Available: http://dl.acm.org/citation.cfm?id=2140458.2140465

[22] H. Saif, M. Fernandez, and H. Alani, "Automatic stopword generation using contextual semantics for sentiment analysis of twitter," in *CEUR Workshop Proceedings*, vol. 1272, 2014.

[23] C. D. Paice, "An evaluation method for stemming algorithms," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994, pp. 42–50.

[24] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[25] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

[26] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Annals of statistics*, vol. 36, no. 6, p. 2605, 2008.

[27] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, 1997, pp. 412–420.

[28] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[29] V. T. Nguyen, H. N. K. Huy, P. T. Tai, and H. A. Hung, "Improving multi-class text classification method combined the svm classifier with oao and ddag strategies," *Journal of Convergence Information Technology*, vol. 10, no. 2, p. 62, 2015.

Fabio Rangel possui Bacharelado em Ciência da Computação pela Universidade Federal do Rio de Janeiro (UFRJ) em 2015. Atualmente é mestrando pelo Programa de Pós-Graduação em Informática da UFRJ (PPGI - UFRJ) sob orientação de Jonice de Oliveira Sampaio. Suas pesquisas estão associadas a Mineração de Dados, Análise de Sentimento em Mídias Sociais Online, Análise de Redes Sociais e Aprendizado de Máquina.

Fabrício Firmino possui graduação em Ciênca da Computação (2010) e mestrado em Informática (2013) pela Universidade Federal do Rio de Janeiro (UFRJ). Atualmente é doutorando sob orientação de Jonice Oliveira Sampaio no Programa de Pós-Graduação em Informática da UFRJ(PPGI-UFRJ). Seus interesses de pesquisas estão relacionadas com mineração de dados e processamento de imagens em redes sociais.

Samara Alves possui graduação em Ciências Atuariais e Estatística pela Universidade Federal do Rio de Janeiro. Atualmente cursa o mestrado em Informática na área de Modelos e Arquiteturas para Sistemas Inteligentes do Programa de Pós Graduação em Informática - Universidade Federal do Rio de Janeiro.

Jonice de Oliveira Sampaio obteve o seu doutorado em 2007 em Engenharia de Sistemas e Computação, ênfase em Banco de Dados, pela COPPE/UFRJ. Na mesma instituição realizou o seu Pós-Doutorado, concluindo-o em 2008. Atualmente é professora adjunta do Departamento de Ciência da Computação da UFRJ, coordenadora do curso de Análise de Suporte à Decisão (habilitação do Bacharelado em Ciências da Matemática e da Terra) e atua no Programa de Pós-Graduação em Informática (PPGI-UFRJ). Em 2013, tornou-se Jovem Cientista do Nosso Estado pela FAPERJ. Suas principais áreas de pesquisa são Gestão do Conhecimento, Análise de Redes Sociais, Big Data e Computação Móvel.

Priscila M. V. Lima é Professora Adjunta no Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro (UFRJ). Ela graduou-se em Computação no IM/UFRJ em 1982. Obteve seu mestrado na COPPE/UFRJ em 1987 e seu doutorado no Imperial College London, Inglaterra, em 2000, ambos em Ciência da Computação. Seus interesses de pesquisa e ensino abrangem inteligência computacional, lógica computacional e algoritmos distribuídos.