Relatório 3 IA: *Machine Learning*

Erick Grilo[1], **Max Fratane**[1], **Vitor Araujo**[1], **Vítor Lourenço**[1]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói, Rio de Janeiro – Brazil

{simas_grilo,mfratane,vitoraraujo,vitorlourenco}@id.uff.br

Resumo.

1. Introdução

O que é pensado pelas pessoas sempre foi uma informação importante para seres humanos para o processo de tomada de decisão. Com o advento da *World Wide Web*, cresceu o acesso à quantidade de opiniões e experiências sobre determinados assuntos que são de pessoas que não conhecemos e nem são profissionais especialistas no assunto. Dessa forma, é possível obter informações de pessoas com os mais variados sentimentos acerca de algum assunto.

Nesse espectro, surge a área de análise de sentimentos (ou mineiração de opiniões), que é responsável por fazer o processamento de linguagem natural, usando táticas de análise textual e linguística computacional a fim de identificar, extrair e estudar opiniões, estados afetivos e informação subjetiva. Dessa forma, é possível extrair opiniões de consumidores acerca de um determinado produto, por exemplo. Tal mineiração é extremamente útil, pois como é visto em [Pang et al. 2008], influencia bastante em tópicos como a aquisição de serviços: a cada 2000 americanos, dentre os leitores de resenhas on-line de restaurantes, hotéis e outros serviços, como viagens, escolas, médicos e cursos, de 73% à 87% dos entrevistados disseram que tais resenhas tiveram uma influência significante na aquisção desses serviços [Zhu and Zhang 2010].

Tal abordagem também é útil para outras finalidades: além da compra de serviços e produtos, as revisões de outros usuários online também são úteis na busca de opiniões políticas (tanto acerca de empresas e organizações quanto acerca de políticos): muitas pessoas buscam atualmente informações de outras acerca de políticos, por exemplo, para confirmar se a opinião dele é condizente com a sua, ou até mesmo buscam na internet opiniões que divergem das suas a fim de enriquecer o debate [Gil de Zúñiga et al. 2009].

Com o advento de plataformas na *web*, tais como blogs, fóruns de discussão, redes *peer-to-peer* e outros tipos de *social media*, tais como o *Facebook* e o *Twitter*. consumidores têm uma quantidade de informação e uma facilidade de expor sua opinião sem precedentes, sejam elas negativas ou positivas. sobre qualquer produto ou serviço. Nesse âmbito, grandes companhias (bancos, restaurantes, agências de viagem, redes de *fast-food* e muitas outras companhinhas dos mais diversos ramos) buscam ler desse ”apelo”informações relevantes para satisfazer as opiniões dos potenciais clientes; em outras palavras, essas opiniões podem exercer uma influência enorme na formação de opiniões de outros usuários, formando a ”lealdade”à marca, o público consumidor, podendo alavancar ou condenar um determinado produto ou até mesmo a imagem de uma empresa [Hoffman 2008].

2. Metodologia de Pesquisa

A metodologia abordada foi dividida em três partes. A primeira parte consiste na comparação da abordagem de redução de dimensionalidade PCA [Jolliffe 2002] e seleção de atributos RFE [Guyon and Elisseeff 2003] aplicados no classificador SVM [Michalski et al. 2013]. A segunda parte trata da seleção de parâmetros utilizando a técnica de Grid Search [Snoek et al. 2012]. Por fim, a terceira parte confere a execução dos classificadores Naive Bayes, SVM, Decision Tree e Random Forest [Michalski et al. 2013] em cima da base de dados selecionada.

A base de dados utilizada é formada por tweets sobre os produtos e serviços fornecidos pela Apple e fornecida pela Carnegie Mellon University[1] [2]. A base foi tratada segundo a abordagem de *Bag of Words* utilizando a ferramenta NLTK[3]. A seleção da base baseou-se na proximidade das informações nela presente com o conhecimento de mundo dos integrantes do grupo.

Os algoritmos supracitados e técnicas foram implementadas conforme a ferramenta de aprendizado de máquina SciKit-Learn[4] e a linguagem utilizada foi Python[5] *release* 3.5.2.

3. Avaliação Experimental

Tabela 1. Matriz de Confusão Binária: *Naïve Bayes*

Atual\Previsto	**positivo**	**negativo**
positivo	142	21
negativo	0	316

Tabela 2. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	1.00	0.87	0.93
negativo	0.94	1.00	0.97
média	0.96	0.96	0.96
acurácia	0.956158663883		

Tabela 3. Matriz de Confusão Binária: *SVM*

Atual\Previsto	**positivo**	**negativo**
positivo	163	0
negativo	0	316

Tabela 4. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	1.00	1.00	1.00
negativo	1.00	1.00	1.00
média	1.00	1.00	1.00
acurácia	1.0		

4. Conclusão

Referências

Gil de Zúñiga, H., Puig-i Abril, E., and Rojas, H. (2009). Weblogs, traditional sources online and political participation: An assessment of how the internet is changing the political environment. *New media & society*, 11(4):553–574.

[1] http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/twitter-sanders-apple2.zip
[2] http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/twitter-sanders-apple3.zip
[3] http://www.nltk.org/
[4] http://scikit-learn.org/stable/
[5] https://www.python.org/

Tabela 5. Matriz de Confusão Binária: *Decision Tree*

Atual\Previsto	**positivo**	**negativo**
positivo	163	0
negativo	0	316

Tabela 6. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	1.00	1.00	1.00
negativo	1.00	1.00	1.00
média	1.00	1.00	1.00
acurácia		1.0	

Tabela 7. Matriz de Confusão Binária: *Random Forest*

Atual\Previsto	**positivo**	**negativo**
positivo	162	1
negativo	3	313

Tabela 8. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.98	0.99	0.99
negativo	1.00	0.99	0.99
média	0.99	0.99	0.99
acurácia		0.991649269311	

Tabela 9. Matriz de Confusão Ternária: *Naïve Bayes*

Atual\Previsto	**positivo**	**negativo**	**neutro**
positivo	16	1	146
negativo	0	99	217
neutro	0	0	509

Tabela 10. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	1.00	0.10	0.18
negativo	0.99	0.31	0.48
neutro	0.58	1.00	0.74
média	0.78	0.63	0.56
acurácia		0.631578947368	

Tabela 11. Matriz de Confusão Ternária: *SVM*

Atual\Previsto	**positivo**	**negativo**	**neutro**
positivo	40	3	120
negativo	0	118	198
neutro	0	2	507

Tabela 12. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	1.00	0.25	0.39
negativo	0.96	0.37	0.54
neutro	0.61	1.00	0.76
média	0.79	0.67	0.63
acurácia		0.673076923077	

Tabela 13. Matriz de Confusão Ternária: *SVM com Grid Search*

Atual\Previsto	**positivo**	**negativo**	**neutro**
positivo	161	0	2
negativo	0	314	2
neutro	3	0	506

Tabela 14. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.98	0.99	0.98
negativo	1.00	0.99	1.00
neutro	0.99	0.99	0.99
média	0.99	0.99	0.99
acurácia		0.992914979757	

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Hoffman, T. (2008). Online reputation management is hot—but is it ethical. *Compu-*

Tabela 15. Matriz de Confusão Ternária: *Decision Tree*

Atual\Previsto	**positivo**	**negativo**	**neutro**
positivo	162	0	1
negativo	1	314	1
neutro	3	2	504

Tabela 16. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.98	0.99	0.98
negativo	0.99	0.99	0.99
neutro	1.00	0.99	0.99
média	0.99	0.99	0.99
acurácia	0.991902834008		

Tabela 17. Matriz de Confusão Ternária: *Random Forest*

Atual\Previsto	**positivo**	**negativo**	**neutro**
positivo	163	0	5
negativo	0	316	7
neutro	3	1	505

Tabela 18. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.96	0.97	0.96
negativo	1.00	0.97	0.98
neutro	0.98	0.99	0.98
média	0.98	0.98	0.98
acurácia	0.97975708502		

Tabela 19. Matriz de Confusão Quartenário: *Naïve Bayes*

Atual\Previsto	**positivo**	**negativo**	**neutro**	**irrelevante**
positivo	127	20	16	0
negativo	2	309	5	0
neutro	25	132	350	2
irrelevante	14	22	23	26

Tabela 20. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.76	0.78	0.77
negativo	0.64	0.98	0.77
neutro	0.89	0.69	0.78
irrelevante	0.93	0.31	0.46
média	0.80	0.76	0.75
acurácia	0.992914979757		

Tabela 21. Matriz de Confusão Quartenário: *SVM*

Atual\Previsto	**positivo**	**negativo**	**neutro**	**irrelevante**
positivo	78	18	67	0
negativo	17	243	56	0
neutro	23	67	417	2
irrelevante	4	8	41	32

Tabela 22. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.64	0.48	0.55
negativo	0.72	0.77	0.75
neutro	0.72	0.82	0.77
irrelevante	0.94	0.38	0.54
média	0.73	0.72	0.71
acurácia	0.71761416589		

Tabela 23. Matriz de Confusão Quartenário: *SVM com Grid Search*

Atual\Previsto	**positivo**	**negativo**	**neutro**	**irrelevante**
positivo	159	0	4	0
negativo	0	312	4	0
neutro	1	1	507	0
irrelevante	5	2	43	35

Tabela 24. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.96	0.98	0.97
negativo	0.99	0.99	0.99
neutro	0.91	1.00	0.95
irrelevante	1.00	0.41	0.58
média	0.95	0.94	0.94
acurácia	0.944082013048		

terworld, February, pages 1–4.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Tabela 25. Matriz de Confusão Quartenário: *Decision Tree*

Atual\Previsto	**positivo**	**negativo**	**neutro**	**irrelevante**
positivo	163	0	0	0
negativo	0	316	0	0
neutro	4	2	503	0
irrelevante	19	8	50	8

Tabela 26. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.88	1.00	0.93
negativo	0.97	1.00	0.98
neutro	0.91	0.99	0.95
irrelevante	1.00	0.09	0.17
média	0.93	0.92	0.89
acurácia	0.922646784716		

Tabela 27. Matriz de Confusão Quartenário: *Random Forest*

Atual\Previsto	**positivo**	**negativo**	**neutro**	**irrelevante**
positivo	163	0	0	0
negativo	0	316	0	0
neutro	0	1	508	0
irrelevante	7	3	58	17

Tabela 28. Medidas da Matriz de Confusão

	precisão	**recall**	**f1-score**
positivo	0.96	1.00	0.98
negativo	0.99	1.00	0.99
neutro	0.90	1.00	0.95
irrelevante	1.00	0.20	0.33
média	0.94	0.94	0.92
acurácia	0.922646784716		

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (2013). *Machine Learning: An Artificial Intelligence Approach*. Springer Publishing Company, Incorporated.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. pages 2960–2968.

Zhu, F. and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148.