7
Bibliografia

ALUÍSIO, S. M. et al. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. **Computational Processing of the Portuguese Language**, v. 2721, p. 194, 2003.

ALVIM, L. G. M. et al. **Sentiment of Financial News:** A Natural Language Processing Approach. Proceedings of the 1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology. Buenos Aires: [s.n.]. 2010. 10-14 Mai 2010.

ANDROUTSOPOULOS, I. et al. **An evaluation of Naive Bayesian anti-spam filtering**. Proceedings of the workshop on Machine Learning in the New Information Age and 11th European Conference on Machine Learning. Barcelona, Spain: ECML 2000. 2000. p. 9-17.

ANDROUTSOPOULOS, I. et al. **An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages**. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information. New York, NY, USA: ACM. 2000. p. 160-167.

ATKINSON, K. GNU Aspell. **GNU Aspell**, 2004. Disponivel em: <http://aspell.net/>. Acesso em: 20 maio 2011.

BERGER, A. L.; PIETRA, S. A. D.; PIETRA, V. J. D. A maximum entropy approach to natural language processing. **Computational Linguistics**, Cambridge, MA, USA, v. 22, n. 1, p. 39-71, Mar 1996. ISSN 0891-2017.

BIRD, S.; LOPER, E. Natural Language Toolkit. **Natural Language Toolkit**, 2011. Disponivel em: <http://www.nltk.org/>. Acesso em: 28 Jul 2011.

CARRERAS, X.; MÀRQUEZ, L. **Boosting Trees for Anti-Spam Email Filtering**. Proceedings of RANLP-2001. Bulgaria: [s.n.]. 2001. p. 58-64.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics**, Cambridge, v. 16, n. 1, p. 22-29, Mar 1990. ISSN 0891-2017.

PUC-Rio - Certificação Digital Nº 0821390/CA

DOMAINTOOLS.COM. Daily DNS Changes and Web Hosting Activity by DailyChanges.com. **DailyChanges.com**, 2011. Disponivel em: <http://www.dailychanges.com/>. Acesso em: 19 Jan 2011.

DRUCKER, H.; WU, D.; VAPNIK, V. N. Support Vector Machines for Spam Categorization. **Neural Networks, IEEE Transactions on**, v. 10, n. 5, p. 1048 -1054, Set 1999. ISSN 1045-9227.

FAN, R.-E. et al. LIBLINEAR: A Library for Large Linear Classification. **The Journal of Machine Learning Research**, v. 9, p. 1871-1874, Jun 2008. ISSN 1532-4435.

FAVRE, B.; HAKKANI, D. Icsiboost. **Icsiboost - Open-source implementation of Boostexter (Adaboost based classifier)**, 2011. Disponivel em: <http://code.google.come/p/icsiboost>. Acesso em: 28 jul. 2011.

GAMMA, E. et al. **Design Patterns:** Elements of Reusable Object-Oriented Software. 1st Edition. ed. [S.l.]: Addison-Wesley Professional, 1994. ISBN 0201633612.

GNU ASPELL. Spell Checker Test Kernel Results. **GNU Aspell**, 2011. Disponivel em: <http://aspell.net/test/cur/>. Acesso em: 16 Jul 2011.

JOACHIMS, T. Text categorization with Support Vector Machines: Learning with many relevant features. In: JOACHIMS, T. **10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings**. Dortmund: Springer Berlin / Heidelberg, v. 1398, 1998. p. 137-142.

KOELING, R. **Chunking with maximum entropy models**. Proceedings of CoNLL-2000 and LLL-2000. Lisbon, Portugual: [s.n.]. 2000.

LAI, C.-C. An empirical study of three machine learning methods for spam filtering. **Know.-Based Syst.**, Amsterdam, v. 20, n. 3, p. 249-254, Abr 2007. ISSN 0950-7051.

LEVENSHTEIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. **Soviet Physics Doklady**, v. 10, p. 707-710, feb 1966.

LEWIS, D. Naive (Bayes) at forty: The independence assumption in information retrieval. In: NÉDELLEC, C. A. R. C. **10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings**. New Jersey: Springer Berlin / Heidelberg, v. 1398, 1998. p. 4-15. ISBN 978-3-540-64417-0.

MINIWATTS MARKETING GROUP. World Internet Usage Statistics News and

PUC-Rio - Certificação Digital Nº 0821390/CA

PUC-Rio - Certificação Digital Nº 0821390/CA

World Population Stats. **Internet World Stats**, 2011. Disponivel em: <http://www.internetworldstats.com/stats.htm>. Acesso em: 19 Jan 2011.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. **Proceeding of the conference on empirical methods in natural language processing (EMNLP, 2002)**, Philadelphia, PA, USA, v. 10, p. 79-86, 6-7 Jul 2002.

RATNAPARKHI, A. **A simple introduction to maximum entropy models for natural language processing**. University of Pennsylvania. Philadelphia. 1997.

SAKKIS, G. et al. **Stacking classifiers for anti-spam filtering of e-mail**. Proceedings of "Empirical Methods in Natural Language Processing". Pittsburgh, PA: EMNLP 2001. 2001. p. 44-50.

SCHAPIRE, R. E.; SINGER, Y. BoosTexter: A boosting-based system for text categorization. **Machine Learning**, v. 2/3, n. 39, p. 135-168, 2000.

SKUT, W.; BRANTS, T. **A maximum entropy partial parser for unrestricted text**. Proceedings of the Sixth Workshop on Very Large Corpora. Montréal, Quebéc: [s.n.]. 1998.

TURNEY, P. D. **Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews**. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics. 2002. p. 417-424.

VAPNIK, V.; LERNER, A. Pattern recognition using generalized portrait method. **Automation and Remote Control**, v. 24, p. 774-780, 1963.

VILELA, P. D. C. S. **Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro**. Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, p. 52. 2011.

WIKIMEDIA FOUNDATION. Cross-site scripting. **Wikipedia:** The Free Encyclopedia, 2011. Disponivel em: <http://en.wikipedia.org/wiki/Cross-site_scripting>. Acesso em: 20 Maio 2011.

YONG-FENG, S.; YAN-PING, Z. Comparison of text categorization algorithms. **Wuhan University Journal of Natural Sciences**, v. 9, n. 5, p. 798-804, 2004.

PUC-Rio - Certificação Digital Nº 0821390/CA

Apêndice A – Resultados da etapa I dos experimentos

As tabelas abaixo listam os resultados obtidos experimentalmente para a etapa I, com os corpora I e II. Os experimentos conduzem combinações de resultados que foram julgadas relevantes, não possuem todas as combinações possíveis.

Algumas observações sobre os resultados são necessárias:

- Ordenação decrescente pelo MCC.
- Extratores:
    - Simples: somente palavras e números.
    - WNL: palavras, números e sinais de pontuação.
    - Corretor: palavras e números, mas com correção das palavras.
    - POS (rápido): com classificação gramatical rápida (ver capítulo 4).
    - POS (NB): com classificação gramatical utilizando o Naive Bayes (ver capítulo 4).
- Colunas:
    - Algoritmo: nome do algoritmo utilizado
    - Param: para o SVM e o Boostexter alguns parâmetros precisam ser configurados. No caso do SVM o parâmetro de margin C e para o Boostexter o número de iterações.

PUC-Rio - Certificação Digital Nº 0821390/CA

- N-grams: número de n-grams
- FE: método para extrair os atributos dos comentários
- Strict: Indica se os comentários com e-mails, links ou código javascript/html foram rejeitados independente do classificador
- RC: Quantidade de comentários reprovados corretamente
- AC: Quantidade de comentários aprovados corretamente
- RI: Quantidade de comentários reprovados incorretamente
- AI: Quantidade de comentários aprovados incorretamente
- Acurácia: métrica de acurácia
- Precisão: métrica de precisão
- Recall: métrica de recall
- F1: F-measure ajustado para o parâmetro β=1 (peso igual para as 2 classes)
- MCC: Coeficiente de correlação

- Resultados para o corpus globo-comments

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=20	3	Simples	Sim	28,540	551,043	22,778	55,044	88.16%	55.61%	34.15%	0.423	**0.375**
SVM	c=50	3	Simples	Sim	28,503	551,017	22,804	55,081	88.15%	55.55%	34.10%	0.423	**0.374**
SVM	c=50	3	WNL	Sim	28,147	551,505	22,316	55,437	88.17%	55.78%	33.68%	0.420	**0.373**
SVM	c=50	3	Correção	Sim	28,781	550,068	23,753	54,803	88.05%	54.79%	34.43%	0.423	**0.372**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=50	3	Simples	Não	27,361	552,976	20,845	56,223	88.28%	56.76%	32.73%	0.415	**0.372**
SVM	c=50	3	POS (rápido)	Sim	28,647	549,909	23,912	54,937	88.01%	54.50%	34.27%	0.421	**0.370**
SVM	c=0.01	2	Simples	Não	19,258	565,498	8,323	64,326	88.95%	69.82%	23.04%	0.346	**0.359**
SVM	c=1	2	Simples	Sim	30,124	543,986	29,835	53,460	87.33%	50.24%	36.04%	0.420	**0.357**
SVM	c=50	2	WNL	Sim	30,163	543,742	30,079	53,421	87.30%	50.07%	36.09%	0.419	**0.356**
SVM	c=1	2	WNL	Sim	29,979	544,151	29,670	53,605	87.33%	50.26%	35.87%	0.419	**0.356**
SVM	c=20	2	WNL	Sim	30,393	542,781	31,040	53,191	87.19%	49.47%	36.36%	0.419	**0.354**
SVM	c=0.01	2	WNL	Sim	20,326	563,241	10,580	63,258	88.77%	65.77%	24.32%	0.355	**0.354**
SVM	c=0.01	2	Simples	Sim	20,288	563,293	10,528	63,296	88.77%	65.84%	24.27%	0.355	**0.354**
SVM	c=1	2	POS (rápido)	Sim	30,257	542,886	30,935	53,327	87.18%	49.45%	36.20%	0.418	**0.353**
SVM	c=0.01	2	Correção	Sim	20,269	563,247	10,574	63,315	88.76%	65.72%	24.25%	0.354	**0.353**
SVM	c=1	2	Simples	Não	29,172	545,418	28,403	54,412	87.40%	50.67%	34.90%	0.413	**0.353**
SVM	c=50	3	POS (NB)	Sim	27,159	549,960	23,861	56,425	87.79%	53.23%	32.49%	0.404	**0.353**
SVM	c=20	2	Simples	Sim	30,665	541,713	32,108	52,919	87.07%	48.85%	36.69%	0.419	**0.352**
SVM	c=1	2	Correção	Sim	30,309	542,444	31,377	53,275	87.12%	49.13%	36.26%	0.417	**0.352**
SVM	c=0.01	2	POS (rápido)	Sim	20,181	563,170	10,651	63,403	88.74%	65.45%	24.14%	0.353	**0.351**
SVM	c=50	2	Correção	Sim	30,878	540,689	33,132	52,706	86.94%	48.24%	36.94%	0.418	**0.350**
SVM	c=50	2	Simples	Sim	30,721	541,065	32,756	52,863	86.98%	48.40%	36.75%	0.418	**0.350**
SVM	c=20	2	Simples	Não	29,402	544,176	29,645	54,182	87.25%	49.79%	35.18%	0.412	**0.350**
SVM	c=50	2	Simples	Não	29,442	543,979	29,842	54,142	87.22%	49.66%	35.22%	0.412	**0.349**
SVM	c=20	2	Correção	Sim	30,686	540,924	32,897	52,898	86.95%	48.26%	36.71%	0.417	**0.349**
Naive Bayes		1	Simples	Sim	26,380	550,920	22,901	57,204	87.81%	53.53%	31.56%	0.397	**0.349**
SVM	c=50	2	POS (rápido)	Sim	30,653	540,887	32,934	52,931	86.94%	48.21%	36.67%	0.417	**0.349**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=20	2	POS (rápido)	Sim	30,583	540,814	33,007	53,001	86.92%	48.09%	36.59%	0.416	**0.348**
SVM	c=1	1	WNL	Sim	24,746	554,103	19,718	58,838	88.05%	55.65%	29.61%	0.387	**0.347**
Naive Bayes		1	WNL	Sim	25,706	551,926	21,895	57,878	87.87%	54.00%	30.75%	0.392	**0.346**
SVM	c=1	1	Simples	Sim	24,953	553,156	20,665	58,631	87.94%	54.70%	29.85%	0.386	**0.344**
SVM	c=1	1	Simples	Não	24,048	554,585	19,236	59,536	88.02%	55.56%	28.77%	0.379	**0.341**
SVM	c=0.01	2	POS (NB)	Sim	19,590	562,723	11,098	63,994	88.58%	63.84%	23.44%	0.343	**0.340**
SVM	c=20	1	WNL	Sim	25,583	550,912	22,909	58,001	87.69%	52.76%	30.61%	0.387	**0.339**
SVM	c=1	2	POS (NB)	Sim	28,967	542,698	31,123	54,617	86.96%	48.21%	34.66%	0.403	**0.338**
SVM	c=20	2	POS (NB)	Sim	29,221	541,674	32,147	54,363	86.84%	47.62%	34.96%	0.403	**0.336**
SVM	c=1	1	POS (rápido)	Sim	23,819	554,175	19,646	59,765	87.92%	54.80%	28.50%	0.375	**0.336**
SVM	c=50	2	POS (NB)	Sim	29,271	541,533	32,288	54,313	86.83%	47.55%	35.02%	0.403	**0.336**
SVM	c=1	1	Correção	Sim	23,803	554,159	19,662	59,781	87.92%	54.76%	28.48%	0.375	**0.336**
SVM	c=50	1	WNL	Sim	26,300	548,142	25,679	57,284	87.38%	50.60%	31.47%	0.388	**0.333**
SVM	c=50	1	POS (rápido)	Sim	24,631	551,081	22,740	58,953	87.57%	52.00%	29.47%	0.376	**0.329**
SVM	c=20	1	Simples	Sim	26,410	546,932	26,889	57,174	87.21%	49.55%	31.60%	0.386	**0.328**
SVM	c=50	1	Simples	Sim	26,440	546,546	27,275	57,144	87.16%	49.22%	31.63%	0.385	**0.327**
SVM	c=50	1	Correção	Sim	24,839	550,276	23,545	58,745	87.48%	51.34%	29.72%	0.376	**0.327**
SVM	c=20	1	Correção	Sim	24,647	550,637	23,184	58,937	87.51%	51.53%	29.49%	0.375	**0.326**
SVM	c=20	1	Simples	Não	25,539	548,063	25,758	58,045	87.25%	49.79%	30.55%	0.379	**0.324**
SVM	c=50	1	Simples	Não	25,736	547,471	26,350	57,848	87.19%	49.41%	30.79%	0.379	**0.323**
SVM	c=1	1	POS (NB)	Sim	27,114	543,612	30,209	56,470	86.81%	47.30%	32.44%	0.385	**0.321**
SVM	c=0.01	1	WNL	Sim	16,626	565,391	8,430	66,958	88.53%	66.36%	19.89%	0.306	**0.321**
SVM	c=0.01	1	Simples	Não	15,103	567,627	6,194	68,481	88.64%	70.92%	18.07%	0.288	**0.320**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=0.01	1	Simples	Sim	16,338	565,495	8,326	67,246	88.50%	66.24%	19.55%	0.302	**0.317**
SVM	c=0.01	1	POS (rápido)	Sim	15,976	565,455	8,366	67,608	88.44%	65.63%	19.11%	0.296	**0.311**
SVM	c=0.01	1	Correção	Sim	15,996	565,411	8,410	67,588	88.44%	65.54%	19.14%	0.296	**0.311**
SVM	c=0.01	1	POS (NB)	Sim	16,300	564,742	9,079	67,284	88.38%	64.23%	19.50%	0.299	**0.310**
SVM	c=50	1	POS (NB)	Sim	28,754	534,235	39,586	54,830	85.64%	42.07%	34.40%	0.379	**0.300**
Boostexter	i=25	1	Simples	Sim	12,337	566,268	7,553	71,247	88.01%	62.03%	14.76%	0.238	**0.261**
Boostexter	i=25	2	Simples	Sim	11,794	566,584	7,237	71,790	87.98%	61.97%	14.11%	0.230	**0.255**
Boostexter	i=25	1	Correção	Sim	11,626	566,505	7,316	71,958	87.94%	61.38%	13.91%	0.227	**0.252**
Boostexter	i=25	2	Correção	Sim	11,559	566,526	7,295	72,025	87.93%	61.31%	13.83%	0.226	**0.251**
Boostexter	i=25	1	POS (rápido)	Sim	11,405	566,627	7,194	72,179	87.93%	61.32%	13.64%	0.223	**0.249**
Boostexter	i=25	2	POS (rápido)	Sim	11,353	566,626	7,195	72,231	87.92%	61.21%	13.58%	0.222	**0.248**
Boostexter	i=25	2	POS (NB)	Sim	10,265	567,137	6,684	73,319	87.83%	60.56%	12.28%	0.204	**0.234**
Boostexter	i=25	1	POS (NB)	Sim	10,184	567,193	6,628	73,400	87.83%	60.58%	12.18%	0.203	**0.233**

- Resultados para o corpus globo-twitter

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=50	3	WNL	Sim	312,937	63,585	39,269	35,418	83.45%	88.85%	89.83%	0.893	**0.524**
SVM	c=0.1	2	WNL	Sim	317,159	60,427	42,427	31,196	83.68%	88.20%	91.04%	0.896	**0.519**
SVM	c=1	2	WNL	Sim	312,067	63,419	39,435	36,288	83.22%	88.78%	89.58%	0.892	**0.518**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=20	3	Simples	Sim	313,858	61,857	40,997	34,497	83.27%	88.45%	90.10%	0.893	**0.514**
SVM	c=10	2	WNL	Sim	310,535	63,796	39,058	37,820	82.96%	88.83%	89.14%	0.890	**0.514**
SVM	c=50	3	Simples	Sim	314,048	61,673	41,181	34,307	83.27%	88.41%	90.15%	0.893	**0.514**
SVM	c=50	2	WNL	Sim	310,488	63,735	39,119	37,867	82.94%	88.81%	89.13%	0.890	**0.513**
SVM	c=5	2	WNL	Sim	310,241	63,874	38,980	38,114	82.91%	88.84%	89.06%	0.889	**0.513**
SVM	c=20	2	WNL	Sim	309,871	63,954	38,900	38,484	82.85%	88.85%	88.95%	0.889	**0.512**
SVM	c=50	3	Simples	Sim	313,468	61,812	41,042	34,887	83.17%	88.42%	89.99%	0.892	**0.512**
SVM	c=50	3	POS (NB)	Sim	315,511	60,497	42,357	32,844	83.33%	88.16%	90.57%	0.894	**0.512**
SVM	c=0.1	2	Simples	Não	318,972	58,158	44,696	29,383	83.58%	87.71%	91.57%	0.896	**0.510**
SVM	c=0.1	2	POS (NB)	Sim	319,061	58,061	44,793	29,294	83.58%	87.69%	91.59%	0.896	**0.510**
SVM	c=1	2	Simples	Não	311,620	62,567	40,287	36,735	82.93%	88.55%	89.45%	0.890	**0.509**
SVM	c=1	2	Simples	Sim	311,901	62,307	40,547	36,454	82.93%	88.50%	89.54%	0.890	**0.508**
SVM	c=0.1	2	Corretor	Sim	318,375	58,273	44,581	29,980	83.48%	87.72%	91.39%	0.895	**0.508**
SVM	c=0.1	2	POS (rápido)	Sim	318,674	58,058	44,796	29,681	83.49%	87.68%	91.48%	0.895	**0.508**
SVM	c=50	3	POS (rápido)	Sim	311,910	62,110	40,744	36,445	82.89%	88.45%	89.54%	0.890	**0.507**
SVM	c=1	2	POS (rápido)	Sim	312,153	61,933	40,921	36,202	82.91%	88.41%	89.61%	0.890	**0.507**
SVM	c=50	3	Corretor	Sim	311,970	61,965	40,889	36,385	82.87%	88.41%	89.56%	0.890	**0.506**
SVM	c=5	2	Simples	Não	309,673	63,127	39,727	38,682	82.62%	88.63%	88.90%	0.888	**0.505**
SVM	c=5	2	Simples	Sim	310,304	62,736	40,118	38,051	82.68%	88.55%	89.08%	0.888	**0.504**
SVM	c=10	2	Simples	Não	309,344	63,261	39,593	39,011	82.58%	88.65%	88.80%	0.887	**0.504**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=1	2	Corretor	Sim	311,160	62,189	40,665	37,195	82.74%	88.44%	89.32%	0.889	**0.504**
SVM	c=0.1	2	Simples	Sim	319,351	57,115	45,739	29,004	83.43%	87.47%	91.67%	0.895	**0.504**
SVM	c=50	2	Simples	Não	308,868	63,432	39,422	39,487	82.51%	88.68%	88.66%	0.887	**0.503**
SVM	c=50	2	Simples	Não	308,663	63,545	39,309	39,692	82.49%	88.70%	88.61%	0.887	**0.503**
SVM	c=20	2	Simples	Não	308,975	63,328	39,526	39,380	82.51%	88.66%	88.70%	0.887	**0.503**
SVM	c=50	2	Simples	Sim	309,160	63,158	39,696	39,195	82.52%	88.62%	88.75%	0.887	**0.502**
SVM	c=10	2	Simples	Sim	309,258	63,073	39,781	39,097	82.52%	88.60%	88.78%	0.887	**0.502**
SVM	c=1	2	POS (NB)	Sim	311,545	61,738	41,116	36,810	82.73%	88.34%	89.43%	0.889	**0.502**
SVM	c=50	2	Simples	Sim	309,193	62,967	39,887	39,162	82.48%	88.57%	88.76%	0.887	**0.501**
SVM	c=5	2	Corretor	Sim	309,596	62,676	40,178	38,759	82.51%	88.51%	88.87%	0.887	**0.501**
SVM	c=5	2	POS (rápido)	Sim	309,639	62,623	40,231	38,716	82.50%	88.50%	88.89%	0.887	**0.500**
SVM	c=20	2	Simples	Sim	308,810	62,986	39,868	39,545	82.40%	88.57%	88.65%	0.886	**0.499**
SVM	c=10	2	POS (rápido)	Sim	309,055	62,786	40,068	39,300	82.41%	88.52%	88.72%	0.886	**0.499**
SVM	c=50	2	POS (rápido)	Sim	308,999	62,790	40,064	39,356	82.40%	88.52%	88.70%	0.886	**0.499**
SVM	c=20	2	POS (rápido)	Sim	308,763	62,894	39,960	39,592	82.37%	88.54%	88.63%	0.886	**0.498**
SVM	c=20	2	Corretor	Sim	308,547	62,975	39,879	39,808	82.34%	88.55%	88.57%	0.886	**0.498**
SVM	c=10	2	Corretor	Sim	308,340	63,030	39,824	40,015	82.31%	88.56%	88.51%	0.885	**0.498**
SVM	c=50	2	Corretor	Sim	307,624	63,289	39,565	40,731	82.20%	88.60%	88.31%	0.885	**0.496**
SVM	c=5	2	POS (NB)	Sim	308,337	62,499	40,355	40,018	82.19%	88.43%	88.51%	0.885	**0.493**
SVM	c=10	2	POS (NB)	Sim	308,438	62,425	40,429	39,917	82.19%	88.41%	88.54%	0.885	**0.493**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=50	2	POS (NB)	Sim	307,922	62,375	40,479	40,433	82.07%	88.38%	88.39%	0.884	**0.490**
SVM	c=20	2	POS (NB)	Sim	307,438	62,520	40,334	40,917	81.99%	88.40%	88.25%	0.883	**0.489**
SVM	c=5	1	WNL	Sim	313,086	57,623	45,231	35,269	82.16%	87.38%	89.88%	0.886	**0.476**
SVM	c=50	1	WNL	Sim	313,375	57,335	45,519	34,980	82.16%	87.32%	89.96%	0.886	**0.475**
SVM	c=20	1	WNL	Sim	313,817	56,863	45,991	34,538	82.15%	87.22%	90.09%	0.886	**0.474**
SVM	c=1	1	WNL	Sim	314,769	55,848	47,006	33,586	82.14%	87.01%	90.36%	0.887	**0.470**
SVM	c=10	1	WNL	Sim	314,101	56,211	46,643	34,254	82.07%	87.07%	90.17%	0.886	**0.470**
SVM	c=1	1	POS (NB)	Sim	312,082	57,384	45,470	36,273	81.88%	87.28%	89.59%	0.884	**0.469**
SVM	c=5	1	Simples	Sim	310,683	58,118	44,736	37,672	81.74%	87.41%	89.19%	0.883	**0.469**
SVM	c=50	1	Simples	Sim	310,292	58,342	44,512	38,063	81.70%	87.45%	89.07%	0.883	**0.469**
SVM	c=0.01	2	WNL	Sim	322,914	50,492	52,362	25,441	82.76%	86.05%	92.70%	0.892	**0.469**
SVM	c=20	1	Simples	Sim	310,776	57,970	44,884	37,579	81.72%	87.38%	89.21%	0.883	**0.468**
SVM	c=5	1	POS (NB)	Sim	309,001	58,965	43,889	39,354	81.55%	87.56%	88.70%	0.881	**0.468**
SVM	c=50	1	Simples	Não	309,840	58,461	44,393	38,515	81.63%	87.47%	88.94%	0.882	**0.468**
SVM	c=10	1	POS (NB)	Sim	308,286	59,291	43,563	40,069	81.46%	87.62%	88.50%	0.881	**0.467**
SVM	c=20	1	POS (NB)	Sim	307,931	59,461	43,393	40,424	81.42%	87.65%	88.40%	0.880	**0.467**
SVM	c=50	1	POS (NB)	Sim	307,727	59,534	43,320	40,628	81.39%	87.66%	88.34%	0.880	**0.467**
SVM	c=10	1	Simples	Não	311,204	57,458	45,396	37,151	81.71%	87.27%	89.34%	0.883	**0.466**
SVM	c=10	1	Simples	Sim	310,684	57,650	45,204	37,671	81.63%	87.30%	89.19%	0.882	**0.465**
SVM	c=20	1	Simples	Não	310,681	57,540	45,314	37,674	81.61%	87.27%	89.19%	0.882	**0.464**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=5	1	Simples	Não	311,251	57,134	45,720	37,104	81.64%	87.19%	89.35%	0.883	**0.463**
Naive Bayes		2	POS (NB)	Sim	310,835	56,759	46,095	37,520	81.47%	87.09%	89.23%	0.881	**0.458**
SVM	c=1	1	Simples	Não	313,711	55,061	47,793	34,644	81.73%	86.78%	90.05%	0.884	**0.458**
SVM	c=1	1	Simples	Sim	313,768	54,974	47,880	34,587	81.72%	86.76%	90.07%	0.884	**0.458**
Naive Bayes		2	WNL	Sim	316,304	53,330	49,524	32,051	81.92%	86.46%	90.80%	0.886	**0.457**
SVM	c=0.1	1	WNL	Sim	318,612	51,700	51,154	29,743	82.07%	86.17%	91.46%	0.887	**0.455**
SVM	c=20	1	POS (rápido)	Sim	311,624	55,683	47,171	36,731	81.41%	86.85%	89.46%	0.881	**0.453**
SVM	c=50	1	Corretor	Sim	313,139	54,654	48,200	35,216	81.51%	86.66%	89.89%	0.882	**0.452**
SVM	c=5	1	Corretor	Sim	313,282	54,492	48,362	35,073	81.51%	86.63%	89.93%	0.882	**0.451**
SVM	c=20	1	Corretor	Sim	312,542	54,907	47,947	35,813	81.44%	86.70%	89.72%	0.882	**0.451**
SVM	c=5	1	POS (rápido)	Sim	313,869	54,122	48,732	34,486	81.56%	86.56%	90.10%	0.883	**0.451**
SVM	c=50	1	POS (rápido)	Sim	313,144	54,521	48,333	35,211	81.48%	86.63%	89.89%	0.882	**0.451**
SVM	c=10	1	Corretor	Sim	313,043	54,526	48,328	35,312	81.46%	86.63%	89.86%	0.882	**0.450**
SVM	c=1	1	POS (rápido)	Sim	314,802	53,456	49,398	33,553	81.62%	86.44%	90.37%	0.884	**0.450**
SVM	c=10	1	POS (rápido)	Sim	312,433	54,842	48,012	35,922	81.40%	86.68%	89.69%	0.882	**0.450**
SVM	c=1	1	Corretor	Sim	314,754	53,459	49,395	33,601	81.61%	86.44%	90.35%	0.884	**0.450**
SVM	c=0.01	2	Simples	Não	325,679	46,206	56,648	22,676	82.42%	85.18%	93.49%	0.891	**0.448**
SVM	c=0.01	2	POS (rápido)	Sim	325,435	46,275	56,579	22,920	82.38%	85.19%	93.42%	0.891	**0.447**
SVM	c=0.01	2	Corretor	Sim	325,397	46,278	56,576	22,958	82.37%	85.19%	93.41%	0.891	**0.447**
SVM	c=0.01	2	Simples	Sim	325,755	45,977	56,877	22,600	82.39%	85.14%	93.51%	0.891	**0.447**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=0.1	1	Simples	Não	320,340	48,998	53,856	28,015	81.86%	85.61%	91.96%	0.887	**0.442**
SVM	c=0.01	2	POS (NB)	Sim	326,971	44,528	58,326	21,384	82.33%	84.86%	93.86%	0.891	**0.441**
SVM	c=0.1	1	Simples	Sim	320,296	48,810	54,044	28,059	81.80%	85.56%	91.95%	0.886	**0.440**
SVM	c=0.1	1	Corretor	Sim	320,310	48,503	54,351	28,045	81.74%	85.49%	91.95%	0.886	**0.437**
SVM	c=0.1	1	POS (NB)	Sim	320,266	48,451	54,403	28,089	81.72%	85.48%	91.94%	0.886	**0.436**
SVM	c=0.1	1	POS (rápido)	Sim	320,070	48,558	54,296	28,285	81.70%	85.50%	91.88%	0.886	**0.436**
SVM	c=0.01	1	WNL	Sim	322,189	46,266	56,588	26,166	81.66%	85.06%	92.49%	0.886	**0.428**
SVM	c=0.001	2	WNL	Sim	326,047	41,602	61,252	22,308	81.48%	84.18%	93.60%	0.886	**0.410**
SVM	c=0.01	1	POS (rápido)	Sim	325,135	41,700	61,154	23,220	81.30%	84.17%	93.33%	0.885	**0.405**
SVM	c=0.01	1	Corretor	Sim	325,235	41,542	61,312	23,120	81.29%	84.14%	93.36%	0.885	**0.404**
SVM	c=0.01	1	Simples	Não	325,567	41,239	61,615	22,788	81.29%	84.09%	93.46%	0.885	**0.403**
SVM	c=0.01	1	Simples	Sim	325,517	41,164	61,690	22,838	81.27%	84.07%	93.44%	0.885	**0.402**
Naive Bayes		2	Simples	Não	313,447	48,393	54,461	34,908	80.19%	85.20%	89.98%	0.875	**0.400**
Naive Bayes		2	Simples	Sim	313,557	48,171	54,683	34,798	80.17%	85.15%	90.01%	0.875	**0.399**
Naive Bayes		2	Corretor	Sim	314,708	47,330	55,524	33,647	80.24%	85.00%	90.34%	0.876	**0.397**
SVM	c=0.01	1	POS (NB)	Sim	326,371	39,795	63,059	21,984	81.15%	83.81%	93.69%	0.885	**0.395**
SVM	c=0.001	1	WNL	Sim	325,812	39,880	62,974	22,543	81.05%	83.80%	93.53%	0.884	**0.392**
SVM	c=0.001	2	POS (rápido)	Sim	329,287	36,774	66,080	19,068	81.13%	83.29%	94.53%	0.886	**0.386**
SVM	c=0.001	2	Corretor	Sim	329,230	36,708	66,146	19,125	81.10%	83.27%	94.51%	0.885	**0.385**
SVM	c=0.001	2	Simples	Sim	329,705	35,993	66,861	18,650	81.05%	83.14%	94.65%	0.885	**0.381**

PUC-Rio - Certificação Digital Nº 0821390/CA

Algoritmo	Param	Ngrams	FE	Strict	RC	AC	RI	AI	Acurácia	Precisão	Recall	F1	**MCC**
SVM	c=0.001	2	Simples	Não	329,716	35,939	66,915	18,639	81.04%	83.13%	94.65%	0.885	**0.381**
SVM	c=0.001	2	POS (NB)	Sim	330,988	33,922	68,932	17,367	80.87%	82.76%	95.01%	0.885	**0.370**
SVM	c=0.001	1	POS (rápido)	Sim	329,072	35,049	67,805	19,283	80.70%	82.92%	94.46%	0.883	**0.368**
SVM	c=0.001	1	Corretor	Sim	329,161	34,903	67,951	19,194	80.69%	82.89%	94.49%	0.883	**0.367**
SVM	c=0.001	1	Simples	Não	329,674	34,324	68,530	18,681	80.67%	82.79%	94.64%	0.883	**0.365**
SVM	c=0.001	1	Simples	Sim	329,710	34,286	68,568	18,645	80.67%	82.78%	94.65%	0.883	**0.365**
Naive Bayes		1	POS (NB)	Sim	335,167	29,519	73,335	13,188	80.82%	82.05%	96.21%	0.886	**0.357**
SVM	c=0.001	1	POS (NB)	Sim	330,388	32,129	70,725	17,967	80.34%	82.37%	94.84%	0.882	**0.348**
Boostexter	i=25	2	WNL	Sim	317,237	36,640	66,214	31,118	78.43%	82.73%	91.07%	0.867	**0.313**
Boostexter	i=25	2	Corretor	Sim	326,327	28,673	74,181	22,028	78.68%	81.48%	93.68%	0.872	**0.286**
Boostexter	i=25	2	Simples	Sim	326,787	28,145	74,709	21,568	78.66%	81.39%	93.81%	0.872	**0.284**
Naive Bayes		1	Simples	Não	336,844	21,148	81,706	11,511	79.34%	80.48%	96.70%	0.878	**0.279**
Naive Bayes		1	WNL	Sim	338,270	20,030	82,824	10,085	79.41%	80.33%	97.10%	0.879	**0.279**
Naive Bayes		1	Simples	Sim	336,877	21,045	81,809	11,478	79.33%	80.46%	96.71%	0.878	**0.278**
Naive Bayes		1	Corretor	Sim	338,428	18,447	84,407	9,927	79.09%	80.04%	97.15%	0.878	**0.261**
Boostexter	i=25	2	POS (NB)	Sim	330,404	23,543	79,311	17,951	78.44%	80.64%	94.85%	0.872	**0.257**

PUC-Rio - Certificação Digital Nº 0821390/CA

Apêndice B – Distribuição dos comentários por grupo

Como o objetivo da etapa II era de fazer a classificação de acordo com os grupos, a tabela abaixo contém a quantidade de mensagens por grupo.

- Corpus globo-comments

categoria	qtde
9658	149,116
5603	81,483
6174	51,741
5604	42,845
16031	38,723
5598	32,465
5601	27,443
5602	24,804
7086	22,412
15605	19,781
5605	18,646
5606	15,647
10407	12,431
7085	11,050
9666	10,669
7084	9,307
16814	8,608
9654	8,234
9982	7,301
5599	6,264
10345	5,952
16029	5,496
6091	4,159
16030	3,991
8524	3,928
17671	3,854
46	3,458
16020	3,411
256	3,399
9101	3,026
9356	2,426
9097	2,346
5600	2,117
10041	1,699
105	1,426
10039	1,415
17082	1,281
144	1,251
16619	1,188
47	1,116
76	1,098
16726	1,092
44	1,039
9772	1,033
10040	1,021
94	790
15528	751
16306	599
109	566
39	558
17816	558
17815	442
15526	426
20	408
101	398
37	390
8	368
15637	360
16052	346
27	327
16108	306
15610	278
42	270
31	250
9662	220

PUC-Rio - Certificação Digital Nº 0821390/CA

categoria	qtde
17084	175
16307	148
18402	144
29	138
17083	121
17396	119
8610	84
17814	84
9099	81
148	78
17397	66
45	65
17812	64
15913	60
119	49
15524	48
9076	31
8334	26
9147	25
15530	22
142	15
16725	14
222	11
9141	9
5	9
16107	6
124	5
164	5
214	4
17856	3
178	2
15700	2
43	2
134	1
16634	1
16021	1
16022	1
146	1
147	1
16618	1

- Corpus globo-twitter

Categoria	Aprovados	Reprovados
Araguaia	2.581	4.106
Bom Dia Brasil	128	416
Caldeirão de Twittadas	2.344	7.722
Fantástico	1.493	44.902
G1	2.176	8.434
G1 Carros	56	1.051
G1 Eleições	889	2.566
G1 Pop e Arte	7	78
G1 RJ	31	240
G1 SP	176	3.804
Globo News	152	1.563
Globo Repórter	476	1.681
Hipertensão	4.597	21.920
Inscrições BBB 11	26.361	86.523
Jornal Hoje	219	1.622
Junto & Misturado	1.414	13.288
Passione	19.110	97.774
Ti-Ti-Ti	39.781	49.575
TV Garagem	799	790
Total	**102.790**	**348.055**

Apêndice C – Parâmetros de uso da API de linha de comando

PUC-Rio - Certificação Digital Nº 0821390/CA

```
usage: test_datasets.py [-h] [--prob-distribution] --moderator
                        {baseline,random,svm,nb,boosting,me-iis,me-gis,me-cg}
                        [--feature {simple,pos-fast,pos-nb,wnl,spell}]
                        [--iter ITER] [--cross CROSS] [--ngrams NGRAMS]
                        [--limit LIMIT] [--strict-protection {0,1}]
                        [--svm SVM] [--specific-learn {0,1}]
                        dataset

Measure a moderator/features accuracy (and others metrics)

positional arguments:
  dataset               a dataset file where each line have a status (can be
                        -1 for rejected and 1 for approved), a category (any
                        string without comma) and a text, always in UTF-8. A
                        column separator is (tab character)

optional arguments:
  -h, --help            show this help message and exit
```

PUC-Rio - Certificação Digital Nº 0821390/CA

```
--prob-distribution   Print probabilities distribution
--moderator {baseline,random,svm,nb,boosting,me-iis,me-gis,me-cg}
                      moderator implementation used to train dataset
--feature {simple,pos-fast,pos-nb,wnl,spell}
                      feature extractor method
--iter ITER           number of iterations for boosting and maximum entropy
                      (me)
--cross CROSS         cross validation
--ngrams NGRAMS       use ngram together feature extractor
--limit LIMIT         use only n random elements in dataset
--strict-protection {0,1}
                      Reject comments that have xss, link or e-mail (without
                      learning)
--svm SVM             parameters only for SVM
--specific-learn {0,1}
                      Use specific learning per category
```