Classificação de Prioridade de *Tweets* utilizando Máquinas de Vetor de Suporte

Vinícius Pazzini[1], **Tiago Schenkel**[1], **Mikael Poetsch**[1] **e Ricardo Matsumura Araujo**[1]

[1]Centro de Desenvolvimento Tecnológico
Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brasil

{vspazzini,tschenkel,mpoetsch,ricardo}@inf.ufpel.edu.br

Resumo. *Este artigo provê resultados iniciais sobre a tarefa de classificação automática de prioridade de* tweets*, como forma de amenizar a sobrecarga de informação sofrida por usuários do* Twitter*. Para tanto, aplicamos Máquinas de Vetor de Suporte a um extenso conjunto de exemplos contendo* tweets *manualmente classificados por nove usuários. Mostramos resultados promissores mesmo com a quantidade limitada de informações textuais presente em* tweets*.*

1. Introdução

O *Twitter* é um site de rede social que permite a troca de mensagens curtas (*tweets*) entre seus usuários de maneira rápida e facilitada através de suas inúmeras formas de acesso (portal *web*, aplicativos para dispositivos móveis, dentre outros) [Twitter 2013]. Ao utilizar a ferramenta, o usuário recebe em sua *timeline* todos os *tweets* postados pelas pessoas que ele está seguindo, em ordem cronológica. Devido à popularização do uso da ferramenta, cada vez mais mensagens são trocadas através do *Twitter*, gerando uma grande quantidade de mensagens presentes na *timeline* dos usuários, frequentemente tornando-se inviável ler todas mensagens.

Um dos problemas desta abordagem é que ela pressupõe que o interesse em seguir uma pessoa implica em interesse em todo conteúdo gerado por essa pessoa. No entanto, usuários geram conteúdos de diferentes tipos que são de interesse variado para outros usuários - e.g. um usuário seguindo um político pode estar interessado em conteúdos de política, mas não quando este faz comentários sobre futebol. Isto potencialmente gera uma grande quantidade de *tweets* que não são do interesse do usuário e que tiram a atenção de mensagens mais importantes ou interessantes [Horn 2010].

Para amenizar essa sobrecarga de informações, propomos neste trabalho a aplicação de algoritmos de aprendizado de máquina [Mitchell 1997] no problema de classificação de prioridade de *tweets*, segundo conceitos individuais de relevância e de importância. Neste artigo, mostramos resultados da aplicação de Máquinas de Vetor de Suporte a uma base de exemplos com *tweets* classificados manualmente por nove usuários.

2. Trabalhos Relacionados

Algoritmos de classificação baseados em técnicas de aprendizado de máquina já são amplamente utilizados na criação de filtros automáticos de conteúdos textuais. Para este fim, abordagens supervisionadas e não-supervisionadas têm sido empregadas em serviços de correio eletrônico e redes sociais virtuais. Dentre estes, destacam-se os filtros de mensagens indesejadas ou *SPAM* [Segaran 2008], a caixa de mensagens prioritárias do *Gmail*

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 210–214,
Fortaleza, CE, Brazil, October 21–23, 2013.

[Aberdeen et al. 2010] e o recurso de notícias principais do *Facebook*. Todas essas aplicações práticas visam amenizar a sobrecarga de informações a que o usuário está sujeito, bem como poupar o tempo do usuário.

Há dois grandes diferenciais na aplicação de técnicas semelhantes ao *Twitter*. Por um lado, a quantidade de informação textual presente em um *tweet* é bastante limitada: 140 caracteres. Em oposição, e-mails frequentemente possuem milhares de caracteres. Por outro lado, o *Twitter* oferece informações sociais sobre os usuários que, potencialmente, podem ser atributos úteis na tarefa de classificação.

3. Metodologia

Para o presente trabalho foi utilizado uma base de dados construída por Schenkel (2011), composta de *tweets* classificados manualmente por nove voluntários de acordo com seus interesses individuais. Através de uma ferramenta *web*, a *timeline* dos voluntários foi coletada e armazenada. Posteriormente, os voluntários rotularam seus *tweets* de acordo com seus interesses individuais, separando-os em três categorias: *importante*, *neutro* e *não-importante* [Schenkel 2011].

Além do texto do *tweet* e do rótulo recebido, outros dados foram coletados, tais como: data de publicação, autor da mensagem, aplicativo de origem, número de *retweets*, número de seguidores do autor da mensagem, dentre outros. O conjunto total de atributos coletados está disponível em [Schenkel 2011].

Para facilitar o uso do algoritmo de Máquinas de Vetor de Suporte (*SVM - Support Vector Machines*) [Pilászy 2008], os *tweets* rotulados como *neutro* e *não-importante* foram agrupados numa mesma categoria. Desta forma, as bases de treinamento e de testes elaboradas foram constituídas por *tweets* rotulados como *importantes* e *não-importantes*. O número total de *tweets* rotulados pelos voluntários e a distribuição de classes são mostrados na Tabela 1.

Tabela 1. Total de *tweets* manualmente rotulados pelos voluntários.

Usuário	Importante		Não-Importante		Total
Usuário 1	156	(19,21%)	656	(80,79%)	812
Usuário 2	53	(24,88%)	160	(75,12%)	213
Usuário 3	408	(31,19%)	900	(68,81%)	1308
Usuário 4	75	(32,89%)	153	(67,11%)	228
Usuário 5	63	(06,45%)	913	(93,55%)	976
Usuário 6	52	(03,11%)	1622	(96,89%)	1674
Usuário 7	78	(28,16%)	199	(71,84%)	277
Usuário 8	234	(28,78%)	579	(71,22%)	813
Usuário 9	106	(37,99%)	173	(62,01%)	279

O conjunto de atributos considerados no processo de classificação supervisionada foram: texto do *tweet*, origem, número de *retweets*, *retweeted* (*flag* que indica se a mensagem foi *retweetada*), nome do autor, localização do autor, número de seguidores do autor, número de amigos do autor, fuso horário, idioma, número de listas em que o autor foi adicionado, indicação se o *tweet* foi favoritado pelo usuário, usuários mencionados e *hashtag's* mencionadas.

O texto do *tweet* foi convertido em atributos numéricos através da técnica TF-IDF (*term frequency–inverse document frequency*), que mede o quão importante é cada palavra em relação a coleção de *tweets* coletada [Manning et al. 2008]. De modo a facilitar a criação do classficador, todos *tweets* rotulados foram utilizados, independentemente de seu idioma, ou seja, *tweets* escritos em inglês foram agrupados com *tweets* escritos em português. A implementação do classificador *SVM* foi feita utilizando-se a biblioteca *LIBSVM*, ferramenta *open-source* que fornece a implementação de Máquinas de Vetor de Suporte em diferentes linguagens de programação [Chang and Lin 2011].

Para relatar os resultados, utilizamos a técnica de validação cruzada com 10 partições (*10-fold cross-validation*) [Rezende 2005]. Nesta abordagem, a base de dados é dividida em 10 grupos com um mesmo número de exemplos rotulados. Em seguida, 9 grupos são utilizados para treinar o classificador e o grupo restante é utilizado para se verificar a taxa de acerto.

4. Resultados

A Figura 1 mostra os resultados do treinamento para cada usuário e cada classe. A média de acerto dos *tweets* corretamente classificados ficou em 70,3% para a categoria *importante* e 72,8% para a categoria *não-importante*.

Observa-se que a *SVM* foi capaz de aprender padrões nos dados, uma vez que os resultados apresentados são melhores que o esperado por decisões aleatórias. O algoritmo foi capaz de aprender mesmo em casos onde o número de *tweets* presentes na categoria *importante* era pequeno e muito inferior ao número de *tweets* presentes na categoria *não-importante* (caso dos dados rotulados pelos usuários 5 e 6).

Ainda que em todos os casos o classificador tenha obtido resultados melhores que decisão aleatória, observou-se que há uma grande disparidade na acurácia para diferentes usuários. Para o usuário 1, obtivemos cerca de 62% de acurácia para conteúdos marcados como importantes, enquanto que para o usuário 8 chegou-se a 87,5%. Mais ainda, a relação entre falsos positivos e falsos negativos varia consideravelmente entre usuários.

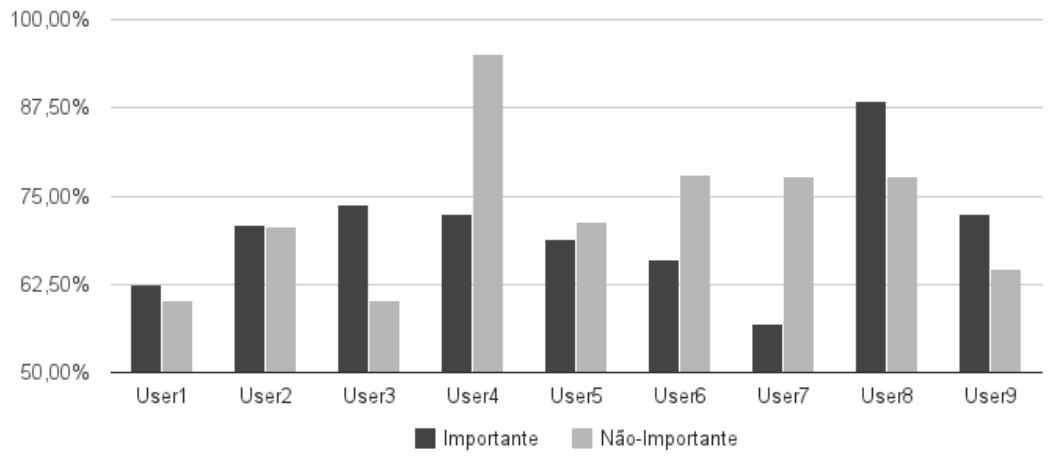

Figura 1. Acurácia do classificador para cada classe e para cada usuário.

Testamos também a acurácia do classificador quando treinado com diferentes atri-

butos. A Figura 2 mostra a acurácia média quando limitamos o número de atributos utilizados. É possível observar que os atributos sociais e textuais são, de certa forma, complementares. Enquanto os atributos sociais são mais eficazes para classificar *tweets* importantes, atributos textuais são mais úteis para classificar *tweets* não-importantes. O uso de todos atributos disponíveis leva a resultados (em média) melhores e mais equilibrados entre as classes.

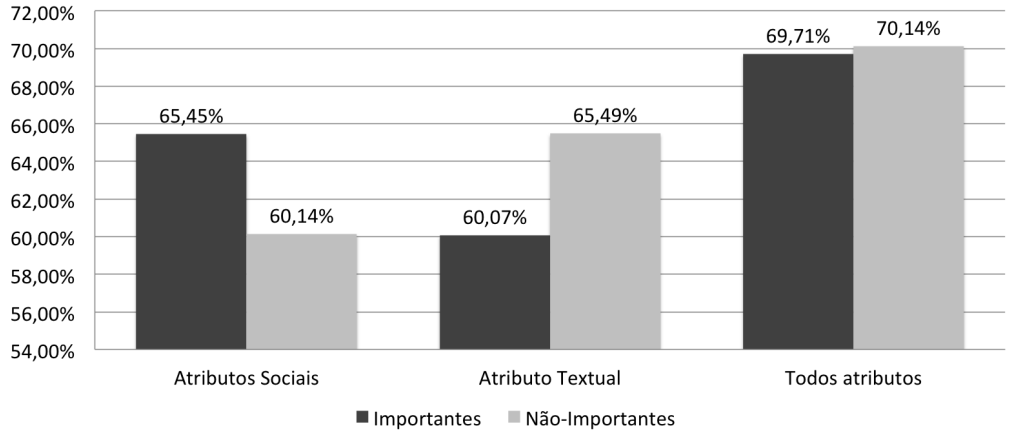

Figura 2. Acurácia média para cada classe utilizando diferentes atributos para treinamento.

5. Conclusões

É possível perceber que existe potencial no emprego de *SVM* para classificar *tweets* de acordo com prioridades fornecidas pelos usuários. Os resultados obtidos até o momento são promissores frente a limitação de um *tweet* conter no máximo 140 caracteres. Observou-se que a utilização de atributos sociais, característica diferencial dos sites de rede social, são úteis para obter boa acurácia, em particular na identificação de conteúdos classificados como importantes.

Compreender o motivo das disparidades observadas de acurácia entre usuários e entre classes é necessário para construir bons classificadores e é um problema que estamos atualmente focando. Também é necessário o teste com diferentes algoritmos de aprendizado de máquina - experimentos iniciais utilizando *Naïve Bayes* [Mitchell 1997] mostraram-se promissores.

O presente trabalho foi realizado com o apoio do UOL através do Programa UOL Bolsa Pesquisa, processo número 20120130151500.

Referências

[Aberdeen et al. 2010] Aberdeen, D., Pacovsky, O., and Slater, A. (2010). The learning behind gmail priority inbox. Technical report, LCCC - NIPS 2010 Workshop on Learning on Cores, Clusters and Clouds., Zurich, Switzerland. Disponível em: http://research.google.com/pubs/archive/36955.pdf Acesso em: 09 de novembro de 2011.

[Chang and Lin 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[Horn 2010] Horn, C. (2010). Analysis and classification of twitter messages. Master's thesis, Graz University of Technology, Graz, Styria, Austria. Disponível em: http://know-center.tugraz.at/wp-content/uploads/2010/12/Master-Thesis-Christopher-Horn.pdf Acesso em: 08 de março de 2013.

[Manning et al. 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, 1st edition.

[Mitchell 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, NY, 1st edition.

[Pilászy 2008] Pilászy, I. (2008). Text categorization and support vector machines. Technical report, Department of Measurement and Information Systems, Budapest University of Technology and Economics. Disponível em: http://conf.uni-obuda.hu//mtn2005/Pilaszy.pdf Acesso em: 14 de abril de 2013.

[Rezende 2005] Rezende, S. O. (2005). *Sistemas Inteligentes*. Manole Ltda., Barueri, SP, 1st edition.

[Schenkel 2011] Schenkel, T. (2011). Além dos filtros sociais: aprendizado de máquina aplicado a personalização de mensagens no twitter. Trabalho acadêmico, Universidade Federal de Pelotas, Pelotas, RS. Disponível em: http://inf.ufpel.edu.br/nopcc/lib/exe/fetch.php?media=monografias:2011:2011-mono-tiago_schenkel.pdf Acesso em: 14 de abril de 2013.

[Segaran 2008] Segaran, T. (2008). *Programando a Inteligência Coletiva*. Alta Books, Rio de Janeiro, RJ, 1st edition.

[Twitter 2013] Twitter (2013). A forma mais rápida e simples de ficar perto de tudo que você gosta. Disponível em: https://twitter.com/about Acesso em: 21 de abril de 2013.