FILTRAGEM DE TWEETS

PAZZINI, Vinícius; POETSCH, Mikael; SCHENKEL, Tiago; ARÁUJO, Ricardo

Universidade Federal de Pelotas
vspazzini,mpoetsch,tschenkel,ricardo {@inf.ufpel.edu.br}.

1 INTRODUÇÃO

O Twitter é uma plataforma social de compartilhamento instantâneo de mensagens, informações e links amplamente utilizado nos dias atuais. Em números extra-oficiais, existem mais de 200 milhões de usuários no Twitter, que compartilham mais de 200 milhões de mensagens por dia (CARLSON, 2011). Conforme CARLSON (2011), que realizou um estudo sobre a relação entre o número de usuários e seguidores, tendo como base 200 milhões de contas, constatou-se que mais de 100 milhões dos usuários seguem entre 2 e 50 pessoas, 15 milhões seguem entre 50 e 500 e apenas 1,5 milhão de usuários seguem mais de 500 pessoas. Com uma quantidade de mensagens gerada, diariamente, tão grande um usuário que não dispõem de muito tempo para verificar sua *timeline* (linha do tempo, onde se organiza e sequencia as postagens conforme o tempo) acaba perdendo informações que poderiam ser importantes ou prioritárias para ele.

O objetivo principal do presente trabalho é testar a utilização de algoritmos de classificação para filtragem. Filtragem é, neste trabalho, o processo de realizar a priorização dos *tweets* de conteúdo do Twitter. O algoritmo classifica e prioriza *tweets* mais importantes conforme os interesses de cada um, possibilitando a filtragem da *timeline*. Um *tweet* é uma mensagem de texto enviada pela rede social Twitter. Utilizam-se as informações contidas nos *tweets* para ensinar ao algoritmo o que é importante e o que não é importante. A principal dificuldade quanto as informações contidas em um *tweet* é o tamanho limitado de caracteres que ele pode conter (140 no máximo).

Filtros inteligentes tem sido aplicados em correio eletronico a bastante tempo com uma boa eficiência. Porém mensagens eletronicas podem conter um quantidade muito grande de caracteres, possibilitando que o filtro trabalhe com um grande número de palavras o que facilita o processo. Diferente de filtragem de tweets, onde a quantidade de tweets é bem mais limitado, fato que dificulta o processo de filtragem, porém outras informações podem ser obtidas de um tweet para auxiliar na filtragem.

O algoritmo de classificação utilizado neste trabalho é um algoritmo de aprendizado de máquina (MITCHELL, 1997). Um algoritmo de aprendizado de máquina utiliza um conjunto de dados (base de treinamento) para aprender os conteúdos que são de preferência do usuário. A partir do conhecimento aprendido ele classifica novos dados, tentando predizer sua importância. Um problema em aprendizado de máquina são os ruídos. Ruídos são informações inconsistentes com a realidade e podem gerar uma grande variação nos resultados de uma classificação. Neste trabalho, propõem-se a utilização de um algoritmo de aprendizagem bayesiano (MITCHELL, 1997) para classificação de *tweets*.

2 METODOLOGIA (MATERIAL E MÉTODOS)

Inicialmente foi utilizada uma base de dados obtida em SCHENKEL(2011), composta por *tweets* classificado manualmente por quatro usuários. Foi utilizado um aplicativo que busca *tweets* da *timeline* de um usuário cadastrado. Foram definidas três possíveis classes para classificação de um tweet, são elas importante, neutro e não-importante. Essa base de dados está disposta como é mostrado na Tabela 1.

Tabela 1 - Distribuição de classificação de tweets

	Total	Importante	Neutro	Não Importante
Usuário1	1674 (39%)	52 (10%)	95 (8%)	1527 (58%)
Usuário2	974 (23%)	63 (12%)	658 (59%)	255 (10%)
Usuário3	813 (19%)	234 (46%)	232 (12%)	347 (13%)
Usuário4	812 (19%)	156 (32%)	137 (12%)	519 (19%)
Total	4275	505	1122	2648

Um *tweet* é composto por um conjunto de atributos: texto, autor, fonte, hora, quantidade de *retweets*, entre outros. O conjunto total de atributos obtidos de um *tweet* esta em SCHENKEL (2011). Os atributos considerados para classificação de um *tweet*, neste trabalho, foram divididos conforme suas características em comum. Foram formados três grupos: textual (atributo texto), geográfico (atributos texto, fonte, localização, fuso horário, língua e data) e social (atributos texto, nome de usuário, número de retweets, número de amigos, número de listas que o autor participa e status count).

Para cada grupo de atributos acima especificado foram feitos testes utilizando as base de dados de um usuário. Para realizar os testes, a base de dados foi divida de forma aleatória em dois conjuntos, um conjunto de treinamento comportando 70% dos dados e um conjunto de teste contendo os 30% restantes. Em cada teste o algoritmo de classificação foi treinado com o conjunto de treinamento e testado no conjunto de teste. Cada classificação feita pelo algoritmo foi comparada à classificação manual feita pelo usuário para validação, gerando assim os resultados apresentados na seção seguinte.

Foram observados os resultados dos diferentes conjuntos de atributos e comparados, buscando qual o conjunto com melhor adequação à classificação dos tweets. Conjuntos que apresentaram resultados muito ruins poderiam ser considerados desnecessários.

3 RESULTADOS E DISCUSSÃO

Inicialmente foram feitos testes utilizando todos os atributos disponíveis, após foram feitos testes utilizando conjuntos reduzidos de atributos com intuito de identificar quais atributos são mais relevantes para cada uma das possíveis classes.

A tabela 2 resume os resultados obtidos nos testes descritos. Cada tabela corresponde a um conjunto de atributos. Cada coluna em uma tabela corresponde a um usuário e o valor nos campos correspondem ao percentual de acerto do algoritmo ao classificar um tweet.

Tabela 2 - Resultados obtidos utilizando todos atributos

Todos atributos				

	Usuário1	Usuário2	Usuário3	Usuário4
Total	56.36%	48.99%	69.98%	47.46%
Importante	22.92%	56.03%	75.2%	37.11%
Não Importante	58.54%	75.05%	77.21%	53.56%
Neutro	39.61%	38.21%	53.85%	36.18%

Observando a tabela 1, é possível observar que os resultados obtidos na tabela 2 são melhores do que se o algoritmo “chutasse” os resultados, baseado na chance base de cada categoria. O usuário3 obteve melhores resultados pois possui uma base de dados mais homogénea, considerando que não há uma diferença significativa quanto aos números de *tweets* classificados em cada categoria.

Tabela 3 - Resultados obtidos utilizando atributos Textual e Menções

	Usuário1	Usuário2	Usuário3	Usuário4
Total	56.55%	49.37%	70.46%	46.79%
Importante	20.67%	57.49%	75.58%	38.39%
Não Importante	58.9%	75.77%	77.26%	52.51%
Neutro	38.61%	38.38%	55.08%	34.77%

Os resultados da tabela assemelha-se aos resultados obtidos na tabela 2, que utiliza todos os atributos na hora de aprendizado e classificação dos *tweets*. Para uma classificação mais precisa são utilizados todos os atributos possíveis, como os resultados se aproximam dos resultados da tabela 2, é possível concluir que os atributos utilizados para a classificação “textual + menções” são os atributos de maior relevância para a classificação de *tweets*.

Tabela 4 - Resultados obtidos utilizando o atributo Textual apenas

	Usuário1	Usuário2	Usuário3	Usuário4
Total	89.54%	66.91%	56.88%	61.21%
Importante	0%	8.93%	53.13%	24.85%
Não Importante	98.04%	31.89%	77.58%	85.28%
Neutro	0.89%	85.97%	29.64%	11.9%

Tabela 5 - Resultados obtidos utilizando atributos sociais apenas

	Usuário1	Usuário2	Usuário3	Usuário4
Total	89.55%	67.14%	57.89%	61.27%
Importante	0%	9.48%	54.74%	25.46%
Não Importante	98.08%	32.67%	78.89%	85.32%
Neutro	1.05%	85.95%	29.49%	11.19%

Tabela 6 - Resultados obtidos utilizando atributos geográficos

	Usuário1	Usuário2	Usuário3	Usuário4
Total	89.57%	67.16%	57.79%	60.48%
Importante	0.17%	8.95%	52.86%	23.18%
Não Importante	98.12%	33.15%	78.02%	85.38%
Neutro	0.61%	85.83%	32.41%	9.01%

Para todos os resultados dos teste das tabelas 4,5 e 6 é possível notar que para cada usuário, a classe com maior taxa de acerto foi aquela que possuía maior quantidade na base de dados de treinamento, como pode ser visto na tabela 1.

4 CONCLUSÕES

A tarefa de classificar um tweet baseado no conjunto de atributos disponíveis mostrou-se um desafio. O fato de cada tweet ser constituído de no máximo 140 caracteres limita a quantidade de informação que se pode trabalhar torna a classificação uma tarefa difícil. Outro ponto problemático é que na base de dados há muito mais tweets rotulados como não importantes em comparação a quantidade de rotulados como importantes e neutros.

A metodologia empregada pelo algoritmo *naive bayes* mostrou-se ineficiente. Observando os resultados da tabela 2, é possível notar que os resultados variam muito de um usuário para outro, tendo acertado de 22% a 75% dos *tweets* importantes, não havendo nenhuma garantia de que o algoritmo conseguiria resultados satisfatórios para um novo usuário.

A utilização do algoritmo para a classificação de um *tweet* utilizando todos os atributos, teoricamente, seria a melhor opção para garantir uma maior precisão na classificação dos *tweets*. Os resultados obtidos utilizando todos os atribuídos assemelha-se aos resultados obtidos utilizando somente os atributos “textual + menções”. Com base nessa observação, é adequado afirmar que a disponibilidade dos atributos “textual + menções” foram os os que mais compensaram a pouca quantidade de texto disponível em um *tweet*.

Como trabalhos futuros destaca-se a utilização de outros algoritmos de aprendizado de máquina, tais como regressão logística, fisher, entre outros.

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

5 REFERÊNCIAS

CARLSON, N. **CHART OF THE DAY**: how many users does twitter really have? Disponível em: http://www.businessinsider.com/chart-of-the-day-howmany-users-does-twitter-really-have-2011-3> Acesso em: 24 de novembro de 2011.

SCHENKEL, Tiago. **Além dos filtros sociais: aprendizado de máquina aplicado a personalização de fluxos de mensagens no Twitter**. 2011. Trabalho acadêmico, Curso de Bacharelado em Ciência da Computação, Universidade Federal de Pelotas, 2011.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw Hill, 1997.