

Emotte: Uma Ferramenta De Análise de Sentimentos para o Twitter

Yuri Malheiros
Universidade Federal da Paraíba
Rio Tinto - Paraíba, Brasil
yuri@dce.ufpb.br

ABSTRACT

Twitter has hundreds of millions users sending messages (tweets) and expressing their opinions about a myriad of subjects, for instance, what they think about a certain product, or if they liked or not a movie, or their reactions during a soccer game. This massive amount of messages carrying opinions about different things could be valuable for business and institutions. Because, it is possible to monitor many people in real time to obtain what they are expressing in their messages in an automatic, fast, and authentic way. In this work, I present Emotte, a tool to analyze sentiment in messages sent to Twitter using machine learning and natural language processing algorithms. The tool monitors the tweets according to queries, classify their sentiments, and display the results using a chart. Thus, someone using Emotte can compare over time the evolution of the opinion of the Twitter users about certain subjects.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous—*web application*; H.3 [Information Storage and Retrieval]: Miscellaneous—*sentiment analysis*

Keywords

sentiment analysis, machine learning, twitter, application

1. INTRODUÇÃO

As redes sociais na Internet se tornaram um importante meio de comunicação, no qual muitos usuários conseguem interagir através do compartilhamento de mensagens textuais, fotos, vídeos, entre outros tipos de dados suportados. A popularidade das redes sociais aumentou de forma impressionante a quantidade de dados gerados nas diversas aplicações desse tipo. O Twitter, uma das redes sociais mais populares atualmente, possui 271 milhões de usuários ativos por mês, que compartilham aproximadamente 500 milhões de *tweets* (mensagens de até 140 caracteres) por dia [1].

Nas redes sociais, é comum os usuários expressarem opiniões sobre diversos assuntos, por exemplo, o que eles acharam do lançamento de um produto, se eles gostaram ou não de um livro, suas reações durante uma partida de futebol, etc. Tal riqueza de informações, em uma quantidade muito grande de mensagens, tem despertado o interesse de pesquisadores e empresas que desejam analisar as respostas dos usuários a estímulos diversos e assim conseguir entender melhor o que pessoas ao redor do planeta estão expressando e sentindo.

Com aplicações que analisam sentimentos de mensagens em redes sociais é possível monitorar em tempo real um grande número de usuários de forma não invasiva, ou seja, os usuários não precisam executar nenhuma atividade adicional ou diferente para que eles tenham suas opiniões analisadas. Além disso, uma pesquisa de opinião através da análise de redes sociais é muito mais rápida e autêntica que pesquisas tradicionais, pois ela pode ser feita automaticamente, não sendo necessário pedir para as pessoas responderem um questionário ou algo semelhante, já que as opiniões são compartilhadas espontaneamente pelos usuários.

O Emotte¹, ferramenta apresentada nesse trabalho, é uma aplicação web que monitora *tweets* e classifica automaticamente os sentimentos das mensagens em três categorias distintas: sentimento positivo, sentimento negativo e neutro. Para isso, o Emotte captura os *tweets* de acordo com consultas cadastradas, classifica os seus sentimentos usando aprendizagem de máquina e, por fim, exibe os resultados em um gráfico comparando diferentes consultas. Com isso, usando a ferramenta, é possível analisar e comparar os sentimentos nas mensagens de diversos usuários do Twitter sobre determinados assuntos de forma rápida, autêntica e automática.

O restante do artigo está organizado da seguinte forma. A seção 2 explica conceitos de análise de sentimentos, sua definição e técnicas existentes. Na seção 3 o Emotte é apresentado detalhadamente, isto é, como a aplicação está organizada, quais são os seus componentes, quais são suas principais funcionalidades e quais são as tecnologias usadas. Na seção 4 são descritos alguns trabalhos relacionados. Por fim, na seção 5 apresentamos as conclusões e são dadas direções para trabalhos futuros.

2. ANÁLISE DE SENTIMENTOS

Análise de sentimentos refere-se à uma forma automática de julgar textos geralmente usando técnicas de processamento

¹<http://emotte.yurimalheiros.com/>

de linguagem natural e aprendizagem de máquina. O seu principal objetivo é mensurar e qualificar sentimentos expressos em uma ou mais frases. A análise de sentimentos também pode ser compreendida como uma forma computacional de tratamento de opinião, sentimento e subjetividade em um texto [7]. Nesse contexto, existem duas técnicas principais para se fazer análise de sentimento: usando classificadores de aprendizagem de máquina supervisionada e usando *lexicons* [10].

Um *lexicon* é um conjunto de palavras e expressões associadas a algumas definições. Sua estrutura lembra um dicionário, no qual uma palavra está associada à sua definição. Mas, no *lexicon* as palavras e expressões geralmente estão associadas a informações de sentimentos (polaridade), classe gramatical, semântica, entre outros. Em uma abordagem de classificação usando um *lexicon*, um documento textual é classificado de acordo com as palavras e expressões do documento avaliado que estão contidas no *lexicon* [11].

Os classificadores de aprendizagem de máquina supervisionada são classificadores capazes de adquirir conhecimento utilizando técnicas computacionais. Eles realizam classificações baseadas em experiências acumuladas anteriormente através de uma etapa de treinamento. Para a análise de sentimento, essa abordagem envolve a construção de classificadores a partir de instâncias, textos ou sentenças previamente rotuladas para classificar novas entradas [8].

3. FUNCIONALIDADES

Para alcançar os objetivos almejado foram definidos os requisitos da aplicação:

1. Capturar *tweets* de acordo com uma consulta;
2. Exibir gráficos comparando os sentimentos de duas consultas;
3. Listar *tweets* capturados;
4. Exibir o resultado da classificação para cada *tweet* capturado.

A seguir são detalhadas cada uma das funcionalidades implementadas de acordo com os requisitos definidos.

3.1 Gráfico de Comparação Entre Consultas

A tela principal do Emotte, que pode ser observada na Figura 1, apresenta um gráfico comparando os sentimentos contidos nos *tweets* de duas diferentes consultas. O título do gráfico mostra quais consultas estão sendo comparadas pelo gráfico, por exemplo, na Figura 1, temos a comparação entre as consultas “Burger King” e “McDonalds”.

No eixo x do gráfico, tem-se as horas no GMT+0, e no eixo y tem-se o nível de sentimento contido nos *tweets*. No eixo y quanto mais alto o ponto, mais sentimentos positivos estavam presentes nos *tweets* observados naquele instante de tempo, por outro lado, quanto mais baixo o ponto, mais sentimentos negativos estavam presentes. As linhas vermelhas correspondem aos sentimentos da consulta que também possui a cor vermelha no título, da mesma forma, as linhas em

azul correspondem aos sentimentos da consulta com a cor azul.

Para conseguir obter um valor numérico no eixo y, as mensagens capturadas são agrupadas em conjuntos de 50 e para cada mensagem positiva soma-se 1 e para cada mensagem negativa subtrai-se 1, por fim, o ponto é adicionado ao gráfico de acordo com a mediana dos tempos em que os *tweets* do conjunto foram enviados.

Abaixo do gráfico existem opções de outras consultas que estão sendo monitoradas pela ferramenta. Para visualizar o gráfico das outras consultas basta clicar no link correspondente.

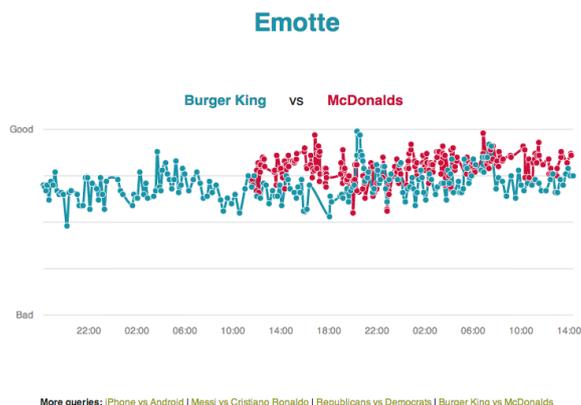


Figura 1: Tela com o gráfico comparando os sentimentos das consultas

3.2 Lista de Tweets Capturados

O gráfico traz uma visão geral dos sentimentos dos *tweets*, mas várias informações podem ser deixadas de lado se não for possível observar individualmente cada mensagem. É exatamente isso que a lista de *tweets* capturados resolve. Ao clicar no nome da consulta no título do gráfico, o usuário é levado para a tela apresentada na Figura 2, que contém todos os *tweets* capturados pela ferramenta para a consulta escolhida. Para cada *tweet*, o usuário do Emotte pode ver quem enviou a mensagem, quando a mensagem foi enviada, qual a mensagem e qual o sentimento (positivo, negativo ou neutro) ela contém. Com isso, o usuário pode analisar mais detalhadamente quem está falando, o que está sendo falado e avaliar o que as classificações realmente refletem sobre a consulta.

4. IMPLEMENTAÇÃO

O Emotte está dividido em três componentes principais:

- Crawler: esse componente é responsável por capturar mensagens através do Twitter;
- Classificador de Sentimentos: cada *tweet* capturado pelo Crawler é classificado pelo Classificador de Sentimentos em uma das três categorias possíveis: positivo, negativo e neutro;
- Apresentação: esse componente é responsável pela interface da aplicação, reagindo as interações do usuário

Emotte



Figura 2: Tela com a lista de tweets capturados para a consulta "mcdonalds"

e apresentando informações gravadas no banco de dados.

A organização dos componentes pode ser vista na Figura 3.

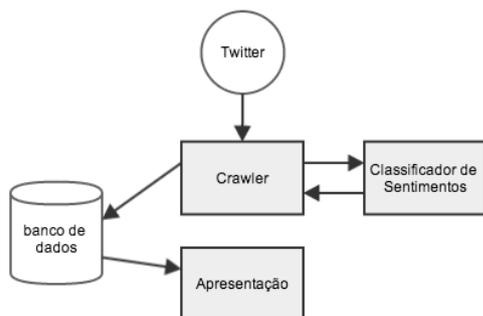


Figura 3: Organização dos componentes do Emotte

4.1 Crawler

O Crawler é o componente responsável por capturar mensagens enviadas para o Twitter através da API² fornecida pela rede social. Ele busca os *tweets* mais recentes enviados de acordo com uma consulta, que pode ser uma palavra ou um conjunto delas. Por exemplo, se for passada a consulta "world cup" para o Crawler, ele busca os *tweets* mais recentes que contenham o texto da consulta.

Após capturar os *tweets*, o Crawler envia cada mensagem para o Classificador de Sentimento para obter a classificação de sentimento correspondente. Em seguida, o Crawler grava o *tweet* com a sua classificação em um banco de dados MongoDB³.

Para o monitoramento dos *tweets* foram definidos no servidor vários *cron jobs* que executam o Crawler a cada cinco

²<http://dev.twitter.com/>

³<http://www.mongodb.org/>

minutos, assim o sistema atualiza seus dados automaticamente quase em tempo real. Tal intervalo foi definido, pois existem limites de uso da API do Twitter e também por que em alguns casos capturar mensagens nessa frequência é suficiente para gravar todos os *tweets* disponíveis. No servidor, cada *cron job* é responsável por executar o Crawler para uma única consulta.

4.2 Classificador de Sentimentos

O Classificador de Sentimentos é o componente responsável por receber uma entrada textual e classificar essa entrada em umas das três categorias: sentimento positivo, sentimento negativo ou neutro. Para isso, o componente usa aprendizagem de máquina supervisionada [6] para efetuar as classificações. Na aplicação foi utilizado o classificador *Passive Agressive* [3] implementado pela biblioteca Scikit-Learn [9], que foi treinado usando o *dataset* Sentiment140 [5].

Esse *dataset* consiste em dois arquivos, um com *tweets* que foram rotulados manualmente e outro com *tweets* rotulados automaticamente de acordo com *emoticons* presentes. Apesar do segundo arquivo ter uma quantidade de dados maior que poderia tornar o classificador melhor, foi constatado, após vários testes, que usar o arquivo rotulado manualmente gerava resultados mais coerentes para o contexto da aplicação. Assim, os dados do arquivo rotulado manualmente foram escolhidos para servir como treinamento.

Antes de ser enviado para o classificador, cada *tweet* é pré-processado. Nessa etapa, todas as letras são transformadas em minúsculas, espaços em branco repetidos são reduzidos para apenas um espaço e nomes de usuários são removidos. Por fim, o texto é transformado em um vetor usando o *Hashing Vectorizer* [13]. Dessa forma, os *tweets* são representado por uma matriz que armazena a ocorrência dos tokens.

4.3 Camada de Apresentação

A Camada de Apresentação é responsável pela interface do Emotte, na qual o usuário pode interagir com a aplicação e receber as informações necessárias de acordo com as funcionalidades do sistema. A interface é estruturada utilizando HTML 5 e o seu estilo é definido usando o *framework* Saas⁴. Algumas bibliotecas JavaScript também são utilizadas, especialmente o JQuery que é responsável no sistema por fazer as requisições AJAX que retornam os *tweets* gravados no banco de dados para serem exibidos na interface.

As duas principais telas do sistema são apresentadas na Figura 1 e Figura 2. A primeira mostra a tela principal com o gráfico comparando duas consultas e, logo embaixo do gráfico, as opções para visualizar as classificações de outras consultas. A segunda mostra a lista com todos os *tweets* capturados para uma determinada consulta, essa tela é exibida ao clicar no nome da consulta que aparece no título do gráfico da tela principal.

5. TRABALHOS RELACIONADOS

Na literatura já foram publicados trabalhos sobre diversas ferramentas que trabalham com análise de sentimentos em mensagens do Twitter. A seguir são apresentados alguns que usam tecnologias semelhantes as do Emotte.

⁴<http://sass-lang.com/>

A ferramenta SPOONS do Netflix [2] usa mensagens dos usuários do Twitter para detectar falhas de disponibilidade no serviço de vídeo. Várias técnicas são utilizadas, entre elas o SPOONS classifica as mensagens em categorias, por exemplo, feliz, neutro e reclamação, para saber se muitos usuários estão reportando problemas. Isto é feito através do uso de classificadores de aprendizagem de máquina supervisionada que são treinados com *tweets* rotulados manualmente por especialistas. Esta abordagem é bastante similar à apresentada, mas possui um escopo mais restrito.

No trabalho de Davidov e colegas [4], ao invés de usar especialistas para rotular manualmente as mensagens de treinamento, foram utilizadas 50 *hashtags* e 15 *emojicons*, estruturas que aparecem naturalmente nos *tweets*, como os possíveis rótulos. Esta é a principal diferença em relação ao Emotte. Dessa forma, o trabalho de classificar as mensagens de treinamento passou a ser automático, possibilitando gerar um conjunto muito maior em um tempo muito menor. Neste trabalho também são utilizados classificadores de aprendizagem de máquina supervisionada.

Wang e colegas [12] descreveram um sistema para análise em tempo real dos sentimentos expressados no Twitter em relação aos candidatos da eleição de 2012 dos Estados Unidos. Um classificador Naive Bayes foi treinado com uma base de dados de aproximadamente 17000 *tweets* (16% positivos, 56% negativos, 18% neutros, 10% incertos), esses *tweets* foram coletados do provedor de dados comercial Gnip Power Track. Baseado nos dados coletados na pesquisa, o classificador conseguiu uma acurácia de 59%.

6. CONCLUSÃO

Nesse trabalho foi apresentada uma ferramenta para análise de sentimentos em mensagens enviadas através do Twitter para conhecer o que os usuários ao redor do planeta estão sentindo sobre determinados assuntos. A principal função da ferramenta é comparar os sentimentos das mensagens entre duas consultas. Com isso, é possível saber, por exemplo, se os usuários do Twitter estão falando melhor sobre uma marca do que de outra.

A aplicação faz o monitoramento dos sentimentos quase que em tempo real. Atualmente ela está configurada para capturar novas mensagens a cada 5 minutos, o que dependendo da consulta é suficiente para capturar todas as mensagens disponíveis pela API. Assim, o Emotte pode trazer uma visão geral para o usuário da evolução dos sentimentos das mensagens sobre uma consulta ao longo do tempo.

Ainda existe muito espaço para a ferramenta evoluir, mas existem dois pontos principais que precisam ser melhorados. O primeiro está relacionado à qualidade do classificador. Um estudo maior precisa ser realizado para que se consiga resultados melhores, pois a análise de sentimentos ainda é um problema difícil para os classificadores atuais. O segundo ponto que precisa ser melhorado é a infraestrutura para suportar muitas consultas, pois o Twitter impõe limitações de uso de sua API e o servidor precisa ser robusto para suportar o processamento e armazenamento de uma quantidade muito grande de mensagens.

7. REFERÊNCIAS

- [1] About twitter, inc. | about. <https://about.twitter.com/company>, July 2014.
- [2] E. Augustine, C. Cushing, A. Dekhtyar, K. McEntee, K. Paterson, and M. Tognetti. Outage detection via real-time social stream analysis: leveraging the power of online complaints. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 13–22, New York, NY, USA, 2012. ACM.
- [3] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.
- [4] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [5] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.
- [6] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.
- [7] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [8] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- [9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [10] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
- [11] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.
- [12] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.
- [13] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120. ACM, 2009.