Extração de Dados do Twitter para aplicação na Análise de Sentimentos

Carlos Augusto F. Filho[1], Fernando M. de Oliveira[1], Ramon de O. M. Lobo[1]
[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ)
CEP – 28.930-000 – Arraial do Cabo – RJ – Brasil

carlos.fernandes@ifrj.edu.br, fernando.oliveira@ifrj.edu.br, ramon.maia.lobo@gmail.com

Abstract. *This work has as objective collect data from the social network Twitter, with the idea to make a Portuguese database to be applied in Sentiment Analysis. Besides the data extracting, a classification algorithm was created based on the emoticons found. This database will be used in future Sentiment Analysis works like the application of others kinds of classification algorithms such as machine learning algorithms.*

1. Introdução

Segundo recente pesquisa realizada em 2013 pelo Comitê Gestor da Internet no Brasil (CGI), o número de usuários de internet no Brasil é de 85,9 milhões. O estudo mostra também que 77% dos usuários de internet utilizam as redes sociais [CGI 2014]. Cada vez mais pessoas postam nas redes sociais suas opiniões, sentimentos e expectativas sobre produtos, empresas e pessoas. Ao pesquisarmos sobre um produto ou pessoa, acessamos constantemente fóruns, blogs de opinião e redes sociais para saber o que as pessoas pensam sobre o referido tema. Esta informação, na maioria das vezes, encontra-se totalmente desestruturada, tornando-se difícil obter uma conclusão positiva ou negativa.

A área de Análise de Sentimentos ou Mineração de Opinião busca descobrir computacionalmente opiniões e seus conceitos relacionados, como: sentimentos, atitudes, avaliações e emoções relacionadas a alguma entidade, tais como: produtos, serviços, organizações, indivíduos, eventos, tópicos e seus atributos. Estas opiniões podem ser na maioria das vezes positivas ou negativas, e em alguns casos neutras [Liu 2012].

O Twitter é uma rede social (*microblogging*), que permite aos usuários enviar e receber atualizações pessoais de outros contatos. Os textos enviados são conhecidos como *tweets* e podem conter no máximo 140 caracteres. A grande maioria das pessoas utiliza a rede social para expressar opiniões e sentimentos. Segundo estudo realizado [Tsytsarau 2010], nos últimos anos pesquisadores têm investido na Análise de Sentimentos no Twitter.

A maioria dos autores que publica na área de análise de sentimentos cria sua própria base de dados e muitas vezes não a disponibiliza. As bases encontradas geralmente estão na língua Inglesa. A técnica mais utilizada e com os melhores resultados na Análise de Sentimentos, tem sido a aplicação de algoritmos de

aprendizado de máquina. Para o treinamento destes algoritmos necessitamos de uma base de dados de *tweets* classificada previamente. Portanto a execução deste trabalho torna-se imprescindível para darmos continuidade às pesquisas relacionadas a esse tema.

2. Proposta

Tendo em vista que o Twitter possui uma API (*Application Programming Interface*) que facilita extração de *tweets* em tempo real, foram feitos alguns testes com a utilização desta API, e verificado seus limites. A versão mais nova da REST API v1.1 permite a execução de 450 requisições a cada 15 minutos. A cada requisição podem ser extraídos 100 *tweets*, totalizando 45.000 a cada 15 minutos. Esse número nos possibilita a criação de uma base robusta e consistente.

Um dos maiores problemas na área de Análise de Sentimentos é a classificação manual da base de dados [B. Liu 2012]. Fica praticamente inviável juntar um grupo de pessoas para classificar milhares de *tweets*. Devido a este problema, foi escolhida a abordagem utilizada por [Go et al. 2009] [Pak 2010] baseado nos *emoticons*. *Tweets* que apresentam *emoticons* positivos ":-)", ":)", "=)", ":D", serão classificados automaticamente como positivos e *tweets* que apresentam *emoticons* negativos ":-(", ":(", "=(", ";(" serão classificados automaticamente como negativos. Apesar de ser uma forma simples de classificação, vários autores utilizaram esta abordagem e conseguiram excelentes resultados. Existem falhas na classificação devido a *tweets* com ironia e sarcasmo. Esse é um dos principais problemas da área de Análise de Sentimentos.

Para extração e classificação estamos desenvolvendo diferentes algoritmos e temos obtido sucesso na maioria deles. As linguagens escolhidas para implementação foram: *Python* e *Ruby*. A REST API fornece mecanismos de filtros que facilitam a extração dos dados. Alguns tipos de filtros: idioma, palavra-chave, data, localização georreferenciada, etc.

Os algoritmos utilizam os *emoticons* como palavras-chave, retornando somente *tweets* com emoticons. Depois esses *tweets* são armazenados em arquivos texto com sua devida classificação. *Tweets* com mais de um *emoticon* são eliminados.

Com a quantidade de requisições que podem ser feitas através da API do Twitter e com os algoritmos que estão sendo implementados para extração e classificação, podemos em pouco tempo ter uma base de dados de tamanho considerável, já classificada como positivo ou negativo, totalmente em Português.

3. Considerações Finais

Este trabalho permitirá a Análise de Sentimentos em redes sociais de diversos assuntos e questões ligadas ao nosso cotidiano, como a opinião das pessoas sobre um produto específico, uma marca ou algum candidato político por exemplo.

Como proposta para trabalhos futuros, temos a análise da acurácia de algoritmos de aprendizado de máquina tais como: *Naive Bayes, Support Vector Machines, Maximum Entropy*, entre outros, para comparação de resultados utilizando a base criada com esse trabalho.

Referências

CGI Comitê Gestor da Internet. www.cetic.br/media/analises/tic-domicilios-2013.pdf. Acessado em 27/10/2014.

CGI Comitê Gestor da Internet. www.cetic.br/media/docs/publicacoes/2/TIC_DOM_EMP_2013_livro_eletronico.pdf pag.179 Acessado em 27/10/2014.

Go A., Richa B. and Lei H. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009): 1-12.

Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. 2012.

Alexander P. and Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC. 2010.

Tsytsarau M. and Palpanas T. (2010). Survey on Mining Subjective Data on the Web. In Proceedings of the 3rd workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Data Mining and Knowledge Discovery.

Twitter API. https://dev.twitter.com/.