Universidade Federal de Pernambuco
Centro de Informática
Graduação em Ciência da Computação

Aplicação de Deep Learning em Análise de Sentimento em Textos de Microblogging

Guilherme Palma Peixoto

Proposta de Trabalho de Graduação

Orientador: Tsang Ing-Ren

Recife
Setembro 2016

Resumo

Uma forma que se tornou popular de compartilhar conteúdo dentro do contexto da Web 2.0 são os sites de microblogging nos quais seus usuários postam seus pensamentos em formatos de textos curtos e sucintos. O site mais popular de microblogging é o Twitter, que limita seus usuários a postarem textos com no máximo 140 caracteres. Esses textos tem um caráter extremamente opinativo, o que ocasionou um interesse da indústria em analisar o que o público tem comentado sobre suas marcas e produtos dentro dessa rede. Esse trabalho tem como propósito o desenvolvimento de uma ferramenta que realiza análise de sentimento a partir de *tweets*[1], utilizando técnicas de Deep Learning para tal. Primeiro, será desenvolvido um módulo de extração e processamento de dados não estruturados do Twitter, com o uso de APIs públicas e técnicas de processamento de linguagem natural. Posteriormente, será realizado o desenvolvimento de um algoritmo de classificação binário de fragmentos de textos com o uso de redes neurais convolucionais para classificação e transformação de palavras em vetores reais. Por fim, será realizado uma análise estatística da performance do algoritmo desenvolvido e será realizado um caso de estudo comparativo com outros algoritmos utilizados dento do contexto da classificação de texto curtos, informais e opinativos.

Palavras-chave: Análise de sentimento, classificação de texto, deep learning, processamento de linguagem natural, Twitter, microblogging, redes sociais, mineração de opinião.

[1] *Tweet* é o termo comumente utilizado para denotar um texto curto de até 140 caracteres compartilhados dentro do Twitter.

Abstract

With the spread of the Web 2.0 usage, it has become a popular practice to share content within microblogging websites, in which its users share their thoughts in short and succinct texts. One of the most popular microblogging website is Twitter, which limits its users to post their posts in texts that can have at most 140 characters. Those short texts are highly informal and it usually express opinions, which led to an interest from the industry to mine those opinions in order to better understand how their brands and products are perceived through the market. This work has as its purpose the development of a framework that performs sentiment analysis in tweets[2] that leverages the use of Deep Learning techniques for such. Firstly, it will be developed a non-structured data extraction module from Twitter (by using its public API) and a pre-processing phase with natural language processing techniques. Then, it will be presented a Deep Learning approach for binary text classification and word embedding with the use of convolutional neural networks. Lastly, it will be shown a statistical analysis of the algorithm performance along with a comparative study of how other more traditional algorithms perform within this short and informal text classification context.

Keywords: sentiment analysis, text classification, deep learning, natural language processing, Twitter, microblogging, social networks, opinion mining.

[2] *Tweet* is the term usually attributed to posts that are shared within the Twitter social network and it represents a short text of at most 140 characters.

Sumário

Introdução

Desde o surgimento da Web 2.0, o crescimento de conteúdo gerado pelos usuários da World Wide Web vem crescendo exponencialmente. Um dos principais mecanismos dessa nova era da informação é a interoperabilidade, de forma que o conteúdo não é gerado apenas atráves de desktops, mas de vários devices externos (principalmente o uso de smartphones) que se encontram conectados à rede. Assim, as pessoas estão conectadas durante a maior parte do seu dia, gerando constantemente conteúdo. Uma das aplicações mais populares dentro da Web 2.0 é o uso de redes sociais, nas quais os usuários podem compartilhar diversas informações, como fotos, vídeos, textos. Junto com o surgimento das novas redes sociais e plataformas nas quais os usuários poderiam publicar as suas opiniões, textos e pensamentos, surgiu uma nova "modalidade" de *blogging* chamada de microblogging, na qual os usuários publicam as suas opiniões em curtos textos.

Dentro do contexto de microblogging, dois sites destacaram-se: o Tumblr[3] e o Twitter[4]. O Twitter, particularmente, alcançou um enorme sucesso: é estimado que sejam postados, em média, 500 milhões de tweets[5] por dia a partir de seus usuários. Como muitos desses *tweets* contém curtas opiniões sobre produtos, marcas e outros sujeitos de análise de interesse, surgiu um grande interesse da indústria a fim de minerar essa enorme quantidade de dados que é gerada diariamente.

Uma das principais aplicações dentro de mineração de opinião é análise de sentimento, que consiste em determinar se o alvo da opinião de um texto tem um caráter positivo ou negativo. Assim, muitas ferramentas e técnicas foram desenvolvidas com o propósito de analisar o sentimento dentro desses textos curtos, usualmente com a aplicação de um algoritmo de classificação binário que rotula o sentimento de um texto como positivo ou negativo. No entanto, como não é possível dar como entrada diretamente uma sequência de caracteres como entrada a um algoritmo de classificação, uma etapa de processamento comum é vetorizar esses textos a fim de produzir um vetor real de tamanho fixo. Esses vetores, porém, conforme o tamanho da base de dados cresce, costumam ser de altíssima dimensionalidade (da ordem de milhões de números cada vetor), de forma que abordagens clássicas, mesmo

[3] https://www.tumblr.com/
[4] https://twitter.com/
[5] Fonte: http://www.internetlivestats.com/twitter-statistics/

que efetivas, terminam tornando-se não escaláveis. Novas abordagens então começaram a surgir para suprir essa necessidade de escalabilidade e velocidade enquanto a acurácia das abordagens tradicionais de classificação fosse mantida.

Com o grande avanço na tecnologia na produção de hardwares cada vez mais eficazes (especialmente memória e GPU), as redes neurais com muitas camadas, que caracterizam as redes encontradas dentro do campo de estudo do *deep learning*, tiveram um interesse retomado pela academia. Apesar do conceito original ter aproximadamente 20 anos de idade, apenas com o avanço do poder computacional que foi possível realizar implementações mais eficazes para a era de *big data* atual.

Deep learning foi rapidamente introduzido no campo de visão computacional e reconhecimento de imagens, mas também encontrou seu caminho em processamento de linguagem natural, onde a sua principal contribuição foi encontrar uma forma de reduzir a dimensionalidade e esparsidade das representações vetoriais de sequências de caracteres. No entanto, apenas adaptar a representação das sequências de caracteres em vetores reais não é suficiente, os algoritmos de classificação também precisam ser adaptados para suportar novas representações. Assim, torna-se importante o estudo de analisar novos algoritmos com técnicas que utilizam o estado-da-arte de performance tanto em termos de precisão da classificação, quanto em tomar o máximo de proveito possível das tecnologias físicas que possuímos hoje.

Objetivos

O objetivo principal desse trabalho é a implementação de um classificador binário que utiliza uma abordagem de Deep Learning voltado para a tarefa de análise de sentimento em cima de textos de caráter curto e informal dentro do contexto de microblogging, que contém uma gramática diferenciada daquelas encontradas em sites que contém opiniões escritas de forma mais tradicional (i.e., mais longa e com a gramática mais formal). Além disso, é incluído a implementação de um módulo que irá realizar a coleta da base de dados para treinamento e teste (com o uso da API pública do Twitter) e uma etapa de pré-processamento utilizando técnicas de processamento de linguagem natural. Por fim, também é objetivado a desenvoltura de um estudo comparativo entre a performance do algoritmo desenvolvido e outros classificadores binários.

Estrutura do Trabalho

O trabalho será divido e desenvolvido de acordo com a seguinte estrutura pretendida:

- **Introdução:** aqui serão introduzidos o tema e a motivação para o trabalho;
- **Conceitos técnicos:** nesse capítulo serão introduzidos alguns conceitos básicos acerca das principais tecnologias e algoritmos utilizados nesse trabalho, não necessariamente restritas ou limitadas a:
    - Técnicas de processamento de linguagem natural utilizadas,
    - Classificadores de aprendizagem de máquina,
    - Deep learning, incluindo: conceitos e visões gerais,redes convolucionais, *word embedding* (mapeamento de palavras para vetores reais de baixa dimensão);
- **Desenvolvimento:** esse capítulo será dividido em três partes principais:
    - Desenvolvimento do módulo de extração de dados do Twitter;
    - Desenvolvimento do módulo de pré-processamento de texto;
    - Desenvolvimento do algoritmo de classificação com o uso de redes convolucionais
- **Avaliação:** será dedicada a analisar a performance do algoritmo proposto de acordo com diversas métricas, incluindo o estudo comparativo com a performance de outros algoritmos aplicados ao conjunto de dados.

Cronograma

Atividades	**Agosto**	**Setembro**	**Outubro**	**Novembro**	**Dezembro**
Formulação da proposta					
Revisão bibliográfica					
Desenvolvimento do módulo de mineração do Twitter					
Desenvolvimento do módulo de pré-processamento de texto					
Desenvolvimento do classificador binário de análise de sentimento					
Análise de performance do algoritmo					
Estudo comparativo					
Preparação da defesa					
Defesa					

Possíveis Avaliadores

Os seguintes professores(as) a seguir são considerados como possíveis avaliadores do trabalho desenvolvido a ser entregue:

- Paulo Salgado Gomes de Mattos Neto (CIn / UFPE)
- Luciano de Andrade Barbosa (CIn / UFPE)
- George Darmiton da Cunha Cavalcanti (CIn / UFPE)

Assinaturas

_____

Tsang-Ing Ren
(Orientador)

_____

Guilherme Palma Peixoto
(Aluno)

Referências

GIMPEL, Kevin et al. **Part-of-speech tagging for twitter: Annotation, features, and experiments**. Proceedings Of The 49th Annual Meeting Of The Association For Computational Linguistics: Human Language Technologies. p. 42-47, 2011.

GO, Alec; HUANG, Lei; BHAYANI, Richa. **Twitter sentiment classification using distant supervision**: CS224N Project Report, Stanford 1. 2009. Disponível em: <http://s3.amazonaws.com/academia.edu.documents/34632156/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1473783321&Signature=E1MnObCowXZCeMdxgQmtZjWtyZU=&response-content-disposition=inline; filename=Twitter_Sentiment_Classification_using_D.pdf>. Acesso em: 07 set. 2016.

GOLDBERG, Yoav; LEVY, Omer. **Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method**. 2014. Disponível em: <https://arxiv.org/pdf/1402.3722v1.pdf>. Acesso em: 07 set. 2016.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge: Mit Press, 2016. Disponível em: <http://www.deeplearningbook.org/>. Acesso em: 09 set. 2016.

HU, Minqing; LIU, Bing. **Mining and summarizing customer reviews**. Acm Sigkdd International Conference On Knowledge Discovery And Data Mining. p. 168-177. 22 ago. 2004.

KIM, Yoon. **Convolutional neural networks for sentence classification**: arXiv preprint arXiv:1408.5882. 2014. Disponível em: <http://arxiv.org/pdf/1408.5882.pdf>. Acesso em: 07 set. 2016.

MIKOLOV, Tomas et al. **Distributed representations of words and phrases and their compositionality**. Advances In Neural Information Processing Systems. 2013.

TANG, Duyu et al. **Coooolll: A deep learning system for Twitter sentiment classification**. International Workshop On Semantic Evaluation. p. 208-212, 2014.