BuscaOpinioes: Searching for Opinions over the Internet

Pedro Paulo Balage Filho and Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos - SP, Brazil
{balage,taspardo}@icmc.usp.br

Abstract. This paper describes the BuscaOpinioes website, a tool for searching for opinions over the internet. Our system uses Google search engine to retrieve reviews from the internet and a lexicon-based sentiment analysis approach to identify opinions in these reviews. A web interface is available to visualize the results as well as some statistics.

Keywords: Sentiment Analysis, Opinion Mining

1 Introduction

We present in this paper a tool for searching for opinions over the internet. Our system, named BuscaOpinioes, is domain independent and it uses natural language processing and sentiment analysis techniques to extract the opinions from the internet pages.

With the advent of Big Data, the information processing technologies have to fit into this scenario of abundant data. The same also happens with sentiment analysis. With the vast amount of information present in product review websites, many users are demanding information on opinions about products and services.

The methodology used in our tool is based on two hypotheses:

i a sample from the most important pages with opinions about a specific topic is likely to have the same positive and negative ratio in relation to all the others;

ii in a collection of reviews, the opinions may be expressed either explicitly (e.g.,"This device sucks") or implicitly (e.g., "This device broke in 2 days!"). We suggest that we are likely to find the same positive and negative ratio in both the explicitly and implicitly opinions.

Although we could not verify these hypotheses with simples experiments, we believe that the reader may empirically observe these hypotheses in the most well-known opinion review websites.

Assuming these two hypotheses, we want to say that our algorithm will be representative of the entire universe of opinions even when it only retrieves a sample of the most important pages and only deals with explicitly opinions.

2 Related work

[2] shows the How Good system, which captures opinions in Twitter for bars and restaurants domain. [4] shows the BestChoice system, a tool for aspect-based sentiment analysis on web texts. [3] presents the OPTIMISM system, which may detect and classify opinions in the politics domain.

For online systems, we have the SCUP[1] web monitoring tool; the Eleitorando[2] portal, specific for sentiment in the elections; and the OpSys[3] tool for opinion mining in stocks.

For Portuguese language, we are not aware of any system that is domain independent or that is able to harvest the whole internet. In this work, we present BuscaOpinioes, a system built to overcome these problems.

3 System Architecture

The architecture of our system comprises five steps:

1. Retrieving of pages containing opinions, using Google search engine;
2. Data pre-processing;
3. Application of lexicon-based sentiment analysis techniques;
4. Sentence ranking;
5. Summarization of the statistics and their exhibition in the website.

In the first step, our system performs a search using the Google Custom Search Engine. This search is built using the user query and the word "reviews". In the Google Custom Search Engine, it is possible to configure which pages the search engine will retrieve in the search. We built a manually list of some well-known product review websites for this purpose. This first step retrieves the first twenty pages using the Google engine.

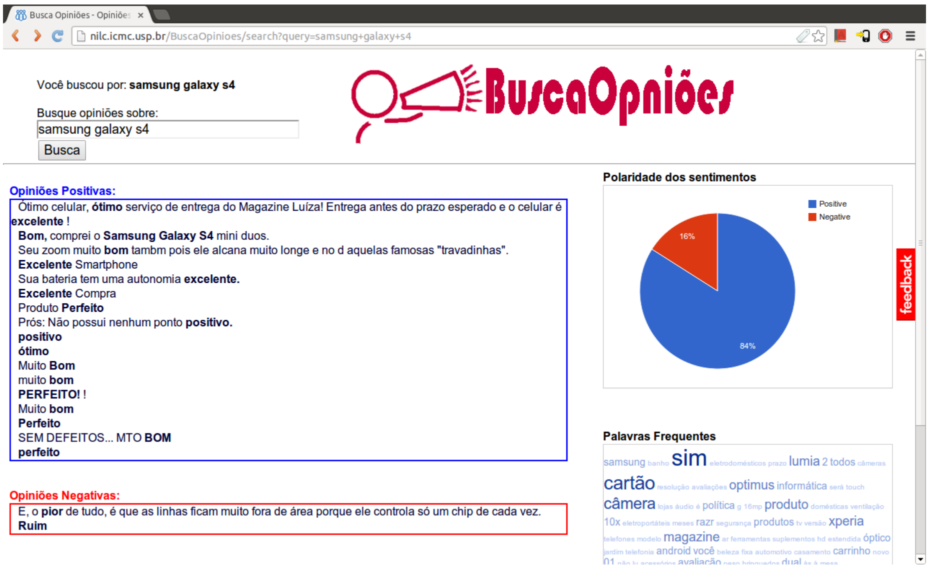
In the second step, the texts obtained from the retrieved pages are sentence segmented, tokenized, and part-of-speech tagged. We used some python NLTK [1] functions for this pre-processing.

The core of our system is in the third step. In this step, we used a manually built lexicon to identify opinions inside the sentences. Our lexicon-based sentiment analysis method uses the polarity of the words in the lexicon to assign a polarity to the sentence. We also built a list of negating words that flips the polarity of the next polar word.

In the fourth step, we rank the best positive and negative sentences based on the third step score. We also look for the presence of the query in order to give some extra weight to these opinions. Finally, the fifth step computes some statistics for the website.

Figure 1 shows the screen dump with the results for the query "Samsung Galaxy S4". Our system is available in the link `http://nilc.icmc.usp.br/BuscaOpinioes`

[1] http://www.scup.com
[2] http://www.eleitorando.com.br
[3] https://www.opsys.com.br


Busca Opiniões - Opiniões
nilc.icmc.usp.br/BuscaOpinioes/search?query=samsung+galaxy+s4
BuscaOpniões
Busca
Polaridade dos sentimentos
Positive
Negative
16%
84%
feedback
Palavras Frequentes
samsung banho sim eletrodomésticos prazo lumia 2 todos câmeras
cartão resolução avaliações optimus informática será touch
câmera lojas áudio é política g 16mp produto domésticas ventilação
10x eletroportáteis meses razr segurança produtos tv versão xperia
telefones modelo magazine ar ferramentas suplementos hd estendida óptico
jardim telefonia android você beleza fixa automotivo casamento carrinho novo
01 não lu acessórios avaliação peso brinquedos dual às à mesa

Você buscou por: **samsung galaxy s4**

Busque opiniões sobre: samsung galaxy s4

Opiniões Positivas:

Ótimo celular, **ótimo** serviço de entrega do Magazine Luíza! Entrega antes do prazo esperado e o celular é **excelente** !
Bom, comprei o **Samsung Galaxy S4** mini duos.
Seu zoom muito **bom** tambm pois ele alcana muito longe e no d aquelas famosas "travadinhas".
Excelente Smartphone
Sua bateria tem uma autonomia **excelente.**
Excelente Compra
Produto **Perfeito**
Prós: Não possui nenhum ponto **positivo.**
positivo
ótimo
Muito **Bom**
muito **bom**
PERFEITO! !
Muito **bom**
Perfeito
SEM DEFEITOS... MTO **BOM**
perfeito

Opiniões Negativas:

E, o **pior** de tudo, é que as linhas ficam muito fora de área porque ele controla só um chip de cada vez.
Ruim

Fig. 1: Screen dump of BuscaOpinioes website

The reader may see two boxes in the interface, a blue one and a red one. They show the positive and negative sentences respectively. For each sentence presented in these boxes, the user may also visualize some information about the page they were originally retrieved from.

The pie chart shows the proportion of positive and negative sentences that were found. Also, we added a list of frequent words used with the query that the user provided.

Due to the architecture of our system, the user may search for any domain and retrive real time information from the whole internet.

Acknowledgments. The authors are grateful to FAPESP for supporting this work.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. 1st edn. O'Reilly Media, Inc. (2009)
2. Fernandes, F.M.d.M.: Um Framework para Análise de Sentimento em Comentários sobre Produtos em Redes Sociais (2010)
3. Silva, M.J., Carvalho, P., Sarmento, L., Magalhães, P., Oliveira, E.: The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence (October 2009) 565–576
4. Silva, N.G.R.d.: BestChoice : Classificação de Sentimento em Ferramentas de Expressão de Opinião (2010)