BlogSet-BR: A Brazilian Portuguese Blog Corpus

[1]Henrique D.P. dos Santos, [2]Vinicius Woloszyn, [1]Renata Vieira
[1]Pontificia Universidade Catolica do Rio Grande do Sul
Av. Ipiranga, 6681, Building 32 - Porto Alegre - Brazil

[2]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

henrique.santos.003@acad.pucrs.br, vwoloszyn@inf.ufrgs.br, renata.vieira@pucrs.br

Abstract

The rich user-generated content found on internet blogs have always attracted the interest of scientific communities for many different purposes, such as from opinion and sentiment mining, information extraction or topic discovery. Nonetheless, an extensive corpora is essential to perform most of Natural Language Processing involved in these tasks. This paper presents BlogSet-BR, an extensive Brazilian Portuguese corpus containing 2.1 billions words extracted from 7.4 millions posts over 808 thousand different Brazilian blogs. Additionally, a survey was conducted with authors to draw a profile of Brazilian bloggers.

Keywords: Blog as Corpus, BlogSet-BR, Brazilian Portuguese Corpus

1. Introduction

Several efforts have been made to build a large corpora based on user-generated content, since they are crucial for many different Natural Language Processing tasks, such as opinion mining, sentiment analysis, topic detection and Age/Gender detection (Burton et al., 2009; Buck et al., 2014; Santos et al., 2017). From all content available on the internet, blogs have been often employed as a main source of user-generated content (Agarwal and Liu, 2008; Agarwal and Liu, 2009; Santos et al., 2012). Nonetheless, there is still a lack of a large semi-structured corpus that also contains author profiles in Brazilian Portuguese. To illustrate this situation, Table 1[1] gives an overview of the initiatives to create a Portuguese corpora based on general content.

Besides Linguateca resources in Table 1, we added other four corpora: the brWac corpus for Brazilian Portuguese built by downloading text from the web (Boos et al., 2014), Buscapé, a Portuguese product reviews corpus extracted from a collaborative review site (Hartmann et al., 2014), the Portuguese Wikipedia dump and BlogSet-BR corpus. The first three corpora listed in the table contains the authorship of the content.

This paper describes BlogSet-BR, a semi-structured large corpus containing author information extracted from Brazilian Portuguese blogs. It contains more than 7.4 millions posts resulting in 2.1 billion words. It is the first Brazilian Portuguese corpus about blogs. In this paper, we also conducted a survey with the authors to create a profile of Brazilian bloggers. The main contributions of this work are:

- The first semi-structured large corpus of posts from Brazilian Portuguese blogs;
- A corpus with authorship, date-time and label information attached to the text;
- Information retrieved from the author profiles, such as gender, age, and educational level.

This work describes the building process of the BlogSet-BR collection and also descriptive statistics about the data. The rest of this paper is organized as follow: first, Section 2. presents related work; in Section 3., we discuss the three phases of creating the BlogSet-BR collection; section 4. provides an overview of the statistics and content on the collected data; section 5. describes the profile of Brazilian bloggers obtained through a survey conducted with authors; section 6. presents the conclusions and suggestions for further work.

2. Background

Extract content from the web is a constant effort by academic and industry researchers. Such data sets allow the accomplishment of many different tasks, such as relevance ranking for online documents and several other machine learning tasks(Woloszyn et al., 2016; Woloszyn et al., 2017). For instance, the Common Crawl project maintains an open repository of web crawl data that can be accessed and analyzed by any research group[2]. This corpus has been used to build language models (Roziewski and Stokowiec, 2016), and to analyze word frequencies (Buck et al., 2014). However, this project does not contain structured information to allow a social network analyses about the authors. This gap is filled by datasets built with blog content, enabling analyses of web data together with user relation and temporal characteristics. The TREC Conference organizers built a blog corpus for research purposes (Macdonald and Ounis, 2006) making 100,649 blogs in the English language available for TREC shared tasks.

The ICWSM 2009 Spinn3r dataset (Burton et al., 2009) is another example of corpora extracted from blogs. It has more than 44 million blog posts for the English language

[1]Adapted from http://www.linguateca.pt/ACDC/

[2]http://commoncrawl.org/

Corpus	Tokens	Words	Sentences	Avg W/S	Authorship
BlogSet-BR	**2,750,700,677**	**2,146,206,009**	**86,803,291**	**24.72**	**Yes**
ptwiki-20170701	390,200,280	341,609,234	79,182,754	4.31	Yes
Buscapé	4,097,905	3,523,417	180,688	19.50	Yes
brWaC	3,207,918,165	2,764,098,344	145,370,673	19.01	No
Corpus Brasileiro	1,175,568,626	990,061,955	45,457,774	21.97	No
NILC/Sao Carlos	42,912,644	32,461,799	1,988,621	16.32	No
ANCIB	1,707,731	1,257,109	83,509	15.05	No
OBras	1,424,014	1,133,302	37,419	30.28	No
ECI-EBR	922,378	723,995	44,381	16.31	No
ReLi	189,577	145,325	8,752	16.60	No
AmostRA-NILC	128,190	98,633	4,931	20.00	No
C-Oral-Brasil	121,092	65,303	9,329	7.00	No
FrasesPB	23,349	19,162	653	29.34	No

Table 1: Brazilian Portuguese Corpora

and has been used in many research works. The work of (Mishne and others, 2005) collected more than 815.494 posts regarding their authors' mood classification, where 10 posts for each of the 40 most frequent moods were annotated. Quan's work (Quan and Ren, 2009) created a fine-grained annotation scheme of emotions in a Chinese corpus containing 1,487 blogs.

To the best of our knowledge, there is no previous work addressing a Portuguese corpora containing an extensible information about the authors based on internet blogs. The BlogSet-BR collection is the first large corpus with information about the profile of the authors for the Portuguese language. Additionally, the dataset contains information that allows analyzing the social network structure. In the next Section, we present the process of developing BlogSet-BR.

3. Corpus Construction

The corpus construction process was performed in three stages: a) selecting suitable blogs to crawl; b) fetching the appropriate content from the Web; and c) organizing the collection into a reusable form.

Selecting the Blogs: The list of blogs used in this dataset was extracted from Santos' work (Santos, 2013) about the Brazilian blogosphere. From this blog list, we selected only the blogs with the 'BR' assigned in the country field of the blog description. Some blogs in the list were discontinued and it was not possible to collect them. In total, 808,973 blogs were selected for the BlogSet-BR collection.

Collecting the Posts: The posts were collected between December 2016 and September 2017 using the Blogspot API[3], which was divided into two tasks: fetching the description of each blog and fetching the posts. Seven crawlers running in parallel were responsible for gathering the contents of the blogs and posts.

Organizing the Collection: Once all crawled data was collected, we had to reorganize it in a format that was easy to use for research purposes. We aimed here to follow the general layout formats from other collections, as this would allow participating groups to reuse existing tools. In the BlogSet-BR collection, we only grouped the blog posts, but the raw data collected from API, in the JSON format, with the blog descriptions and posts content is also available. The BlogSet-BR contains the following fields: blog ID, post ID, author ID, author name, title, published date, content, tags, and number of comments, in the comma-separated values (CSV) format.

[3]https://developers.google.com/blogger/

Figure 1: Top 50 Tag Cloud (translated)

4. Dataset Description

In this section, we show a descriptive statistics about the data and an overview of the BlogSet-BR collection. From all 808 thousand blogs, 7.4 millions posts were collected. They were written by several authors addressing many different subjects. When compressed, the collection reaches 4.7 GB of data.

In the whole dataset, there are more than 57 millions tags. The most frequent tag is 'news', followed by 'tips', 'love', 'music', 'fashion', and 'movies'. Figure 1 shows the top 50 most frequent tags. These tags show the vast amount of content that can be found in BlogSet-BR for text mining and information extraction research.

4.1. Dates and Times

The date and time associated to each post is important to analyze spread of information in social networks (Lerman and Ghosh, 2010; Adar and Adamic, 2005). Figure 2 shows

the distribution of the dates of the collected posts after the year 2003. The posts before 2003 represent only 0.005% of the corpus.

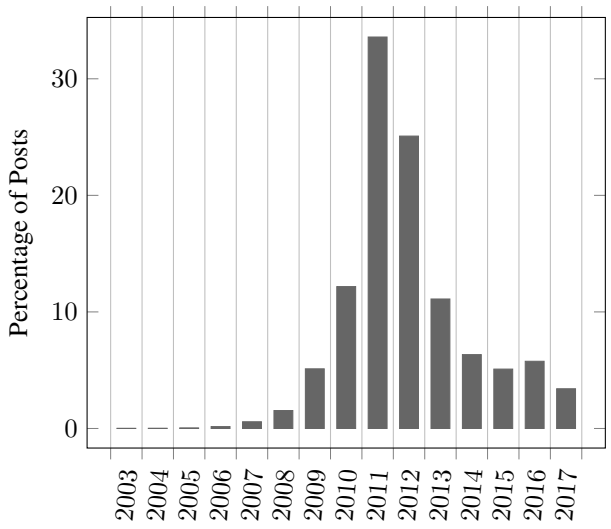

Figure 2: Posts Year Distribution

The posts start decreasing in Blogspot platform after 2011. This behavior could represent the migration of users to Twitter and Facebook, both grow exponentially in Brazil in the same period (Yokoyama and Sekiguchi, 2014).

5. Bloggers Profile

A survey[4] was conducted with 4,332 Brazilian authors from the Blogspot platform to establish a profile of bloggers who use the platform. The requested pieces of information were:

- Age
- Gender
- Topic of interest
- Pageviews
- Update frequency
- Educational level

The majority of authors are male (61%). In addition, almost every user finished high school (92%) and most of them add new a new post at least once a week (85%). Figure 3 shows the age distribution on the collected survey.
Regarding frequency, 32% create new content every day and 53% publish every week. Most users create their own content (86%), giving their own opinion about mundane facts.
Table 2 shows the distribution of topics, which vary: books, fashion, technology, education, politics, movies, and arts. In this question, users were allowed to select multiple choices.
In Figure 4 we show an overview of answers by Brazilian bloggers regarding their educational level. This was not a mandatory question, but 97% of all those who answered

[4]Survey conducted in 2012. Not every author in the survey is in the corpus BlogSet-BR, collected in 2017.

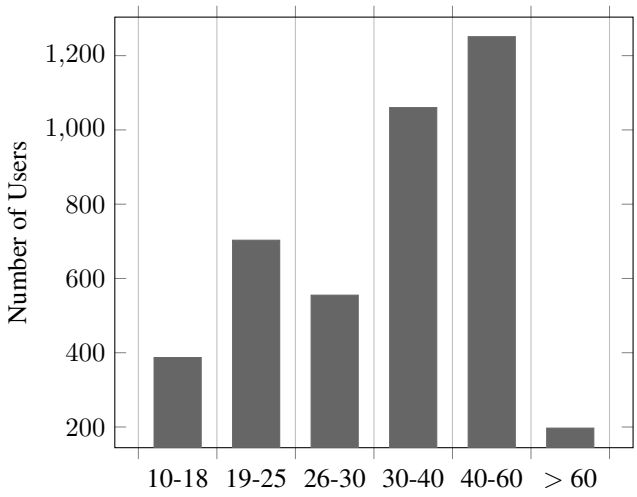

Figure 3: User's Age Distribution

Topic	% of Users
Arts	29%
Education	28%
Literature / Books	28%
Music	27%
Politics	24%
Philosophy	19%
Friendship	17%
Movies	17%
Health	17%
Technology	16%

Table 2: Distribution of the Top 10 Topics selected by users

finished high school and the majority of them are at the university.
Considering social media, most interviewees use Facebook (94%) and Twitter (67%). This rich information about age, gender, and educational level of Brazilian bloggers could be employed in different tasks, such as writing style de-

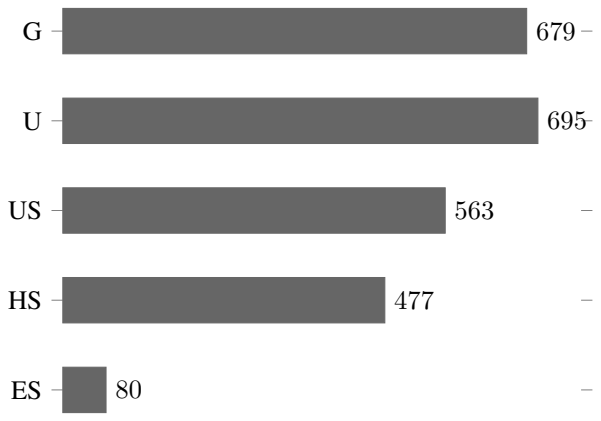

Figure 4: User's Educational Level: G (Graduated), U(Undergraduated), US (Undergraduated Student), HS (High School), ES (Elementary School)

tection and readability assessment (Herring and Paolillo, 2006; Argamon et al., 2007).

6. Conclusion and Further Work

In this paper, we described the motivations, details, and building process of BlogSet-BR. Additionally, we conducted a survey with authors to create their profile, which resulted in a rich description of Brazilian bloggers.
This corpus is the biggest corpus with authorship information for the Brazilian Portuguese language. These types of dataset are useful for topic detection and other NLP tasks.
All the content of this work is available on the web[5]. It is possible to download the raw data in the JSON format, only the Brazilian blog posts in the CSV format, and the survey with 4 thousand users in the XLS format.

Acknowledgments

This work was partially supported by the CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) Foundation (Brazil), PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul), and UFRGS (Universidade Federal do Rio Grande do Sul).

7. Bibliographical References

Adar, E. and Adamic, L. A. (2005). Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence*, pages 207–214. IEEE Computer Society.

Agarwal, N. and Liu, H. (2008). Blogosphere: research issues, tools, and applications. *ACM SIGKDD Explorations Newsletter*, 10(1):18–31.

Agarwal, N. and Liu, H. (2009). Modeling and data mining in blogosphere. *Synthesis lectures on data mining and knowledge discovery*, 1(1):1–109.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *LREC*, volume 2, page 4.

Burton, K., Java, A., and Soboroff, I. (2009). The icwsm 2009 spinn3r dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*. AAAI.

Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Lerman, K. and Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97.

Macdonald, C. and Ounis, I. (2006). The trec blogs06 collection: Creating and analysing a blog test collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, 1:3–1.

Mishne, G. et al. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, pages 321–327. Citeseer.

Quan, C. and Ren, F. (2009). Construction of a blog emotion corpus for chinese emotional expression analysis. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1446–1454. Association for Computational Linguistics.

Roziewski, S. and Stokowiec, W. (2016). Languagecrawl: A generic tool for building language models upon common-crawl. In *LREC*.

Santos, R. L., Macdonald, C., McCreadie, R., Ounis, I., Soboroff, I., et al. (2012). Information retrieval on the blogosphere. *Foundations and Trends® in Information Retrieval*, 6(1):1–125.

Santos, H. D. P., Woloszyn, V., and Vieira, R. (2017). Portuguese personal story analysis and detection in blogs. In *Proceedings of the International Conference on Web Intelligence*, pages 709–715. ACM.

Santos, H. D. P. (2013). Identificação de autoridades em tópicos na blogosfera brasileira usando comentários como relacionamento. Master's thesis, Federal University of Rio Grande do Sul.

Woloszyn, V., Santos, H. D. P., and Wives, L. K. (2016). The influence of readability aspects on the user's perception of helpfulness of online reviews. *Revista de Sistemas de Informação da FSMA*.

Woloszyn, V., Santos, H. D. P., Wives, L. K., and Becker, K. (2017). Mrr: an unsupervised algorithm to rank reviews by relevance. In *Proceedings of the International Conference on Web Intelligence*, pages 877–883. ACM.

Yokoyama, M. H. and Sekiguchi, T. (2014). The use of social network sites in the workplace: A case study in brazilian companies. *Brazilian Business Review*, 11(2):87.

8. Language Resource References

Boos, R., Prestes, K., Villavicencio, A., and Padró, M. (2014). brwac: a wacky corpus for brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 201–206. Springer.

Hartmann, N., Avanço, L., Balage Filho, P. P., Duran, M. S., Nunes, M. d. G. V., Pardo, T. A. S., Aluísio, S. M., et al. (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *LREC*, pages 3865–3871.

[5] http://www.inf.pucrs.br/linatural/blogset-br