UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

BEST SPORTS: A PORTUGUESE COLLECTION OF DOCUMENTS FOR SEMANTICS-CONCERNED TEXT MINING RESEARCH

ROBERTA AKEMI SINOARA
SOLANGE OLIVEIRA REZENDE

Nº 424

RELATÓRIOS TÉCNICOS

São Carlos – SP
Abr./2018

BEST sports: a Portuguese collection of documents for semantics-concerned text mining research

Roberta Akemi Sinoara

Solange Oliveira Rezende

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
Caixa Postal 668
13560-970, São Carlos, SP
e-mail: rsinoara@usp.br, solange@icmc.usp.br

Abstract:

The availability of labeled text collections is a common need in the text mining research community. These collections are used for both learning and evaluating text mining models. In this technical report, we present the *BEST sports* collection. This collection of documents written in Portuguese was collected, prepared, and provided to be used as benchmarking collection in text mining research. Considering real application scenarios, we created four datasets, which correspond to problems of different semantic complexity levels. The use of different datasets of the same text collection allows the evaluation of text mining methods at different levels of semantic complexity.

Keywords: Text mining, Benchmarking Text Collection, Datasets, Semantics.

April/2018

Contents

1. Introduction

A huge amount of text data is created daily, either in social networks and the Web or inside organizations. Text mining techniques have become essential for supporting knowledge discovery as the volume and variety of digital text documents have increased. Text sources, as well as text mining applications, are varied. Among the text mining applications, we can mention automatic text classification, text clustering and topic hierarchy building, sentiment analysis, information extraction, and recovery of documents and multimedia data.

Despite the text mining task, a common need in the text mining research community is the availability of labeled text collections. These collections are used for learning a model and for evaluating it (Aggarwal, 2014; Aggarwal & Zhai, 2012). In automatic text classification process, for instance, supervised machine learning algorithms use labeled data (documents with known class labels) to learn a classification model (classifier). These classification models are also evaluated based on labeled data, comparing the classifier answers for new instances to their true known labels. Besides, the evaluation of unsupervised learning methods is often done with the use of labeled data (a gold standard solution).

In this technical report, we present the *BEST sports* collection, which was prepared for supporting text mining research. *BEST sports* collection is composed of Portuguese news articles about sports and it is organized into four datasets. Each dataset may be seen as an independent gold standard, related to a specific organization objective. We construct the datasets to simulate real text mining application scenarios, in which different users or situations require different organizations (or classifications) for the same text collection. Moreover, having different versions of the same text collection allows the evaluation of text mining methods at different levels of semantic complexity. *BEST sports* documents are available at `http://sites.labic.icmc.usp.br/rsinoara/bestsports`.

The remainder of this technical report is organized as follows. Section 2 describes the *BEST sports* collection. Section 3 presents an analysis of *BEST sports* datasets, considering their semantic complexity. Section 4 presents the labeling process conducted to build the semantic organization datasets.

2. *BEST sports* collection

BEST sports is a collection of sports news, written in Portuguese. The documents were extracted from *BEST sports* website[1], whose main focus is news and results of the Olympic Games and the most important world championships of several sports. The *Best sports* collection contains 881 short news articles (from 383 to 2779 characters), referring to 66 sports or sporting events. The articles were published from August 1999 to August

[1] BEST sports archive: `http://bestsports.com.br/db/notarqhome.php`

2008. The largest amount of documents is from the year 2004, mainly due to Athens Summer Olympics. Each of the 881 articles has a class label, according to the website categorization, and corresponds to either a specific sport or a certain sporting event. The first group of class labels is composed of articles that report games or competition results of the specific sport, for example, an article could report the podium of the first round of Formula 1 World Championship. The second group of class labels is composed of articles that report some fact related to a sporting event in general, as an example, an article could report an Olympic Games Opening Ceremony. Taking into account these class labels, the dataset is unbalanced. Table 1 presents the class distribution of *BEST sports* collection. The top 10 class labels correspond to 502 articles, i.e., more than one half of the total number of documents.

Four datasets are available for *BEST sports* collection. The first dataset, named *BS-full*, is composed of 881 documents organized into 66 classes of the website categorization. The other three datasets are composed of documents from the four larger classes: *Fórmula 1* (Formula 1), *Motovelocidade* (MotoGP), *Futebol* (Soccer), and *Tênis* (Tennis). This subset, named *BEST sports - Top4*, has 283 documents. The first *BEST sports - Top4* dataset corresponds to the website classification. The second and third datasets are related to the performance of Brazilian athletes. In order to create these two datasets, the documents were manually labeled considering this different point of view. Each one of the 283 documents received one of the four possible labels: BRvenceu (*"Brazilian won"*), BRnaoVenceu (*"Brazilian did not win"*), BRnaoCitado (*"No Brazilian mentioned"*) or NaoDefinido (*"Not defined"*)[2]. Details of the labeling process are presented in Section 4.

Thus, the four datasets of the *BEST sports* collection are the following.

1. *BS-full*: categorization by sport;
2. *BS-topic*: categorization by sport;
3. *BS-semantic*: categorization by the performance of Brazilian athletes;
4. *BS-topic-semantic*: categorization by both sport and athletes' performance.

Table 2 presents some characteristics of these four datasets. An analysis of their semantic complexity level is presented in next section.

[2] The label "Not defined" refers to documents that do not report the results of a competition or report both Brazilian victory and defeat.

Table 1: Class distribution of *BEST sports* collection

Class	Label	# Docs
Formula 1	ESP_FORMULA_1	91
Soccer	ESP_FUTEBOL	68
MotoGP	ESP_MOTOVELOCIDADE	64
Tennis	ESP_TENIS	60
Volleyball	ESP_VOLEIBOL	52
Athletics	ESP_ATLETISMO	42
Beach Volleyball	ESP_VOLEI_DE_PRAIA	33
Summer Olympic Games	CMP_JOGOS_OLIMPICOS_VERAO	32
Swimming	ESP_NATACAO	32
Sailing	ESP_IATISMO	28
Basketball	ESP_BASQUETE	22
Gymnastics	ESP_GINASTICA	18
Handball	ESP_HANDEBOL	17
Pan American Games	CMP2_JOGOS_PANAMERICANOS	16
Judo	ESP_JUDO	16
Mountain Biking	ESP_MOUNTAIN_BIKE	16
Motocross	ESP_MOTOCROSS	14
Shooting	ESP_TIRO	14
Equestrian	ESP_HIPISMO	12
Diving	ESP_SALTOS_ORNAMENTAIS	12
Formula 3000	Formula_3000	12
Baseball	ESP_BEISEBOL	11
Cycling	ESP_CICLISMO	11
Rally	ESP_RALI	11
Triathlon	ESP_TRIATLO	11
Cross-Country Rally	ESP_RALI_XC_CARROS	9
Cross-Country Rally Bikes	ESP_RALI_XC_MOTOS	8
Rowing	ESP_REMO	8
TaeKwonDo	ESP_TAEKWONDO	8
Table Tennis	ESP_TENIS_MESA	8
Boxing	ESP_BOXE	7
Wrestling	ESP_LUTA	7
IndyCar	IRL	7
Summer Paralympic Games	CMP2_JOGOS_PARAOLIMPICOS_VERAO	6
Fencing	ESP_ESGRIMA	6
Powerboating	ESP_MOTONAUTICA	6
Supercross	ESP_SUPERCROSS	6
Enduro	ESP_ENDURO	5
Rugby	ESP_RUGBI	5
Archery	ESP_TIRO_ARCO	5
Trial	ESP_TRIAL	5
Badminton	ESP_BADMINTON	4
Canoeing	ESP_CANOAGEM	4
Futsal	ESP_FUTSAL	4
Weightlifting	ESP_LEVANTAM_PESO	4
Champ Car	CART	4
Duathlon	ESP_DUATLO	3
Karate	ESP_KARATE	3
Karting	ESP_KART	3
Synchronised Swimming	ESP_NADO_SINCRONIZADO	3
Modern Pentathlon	ESP_PENTATLO_MODERNO	3
Surfing	ESP_SURF	3
Winter Olympic Games	CMP_JOGOS_OLIMPICOS_INVERNO	2
BMX Cycling	ESP_BICICROSS	2
Skiing	ESP_ESQUI	2
Field Hockey	ESP_HOQUEI	2
Water polo	ESP_POLO_AQUATICO	2
Squash	ESP_SQUASH	2
Superbike	ESP_SUPERBIKE	2
Chess	Xadrez	2
Winter Paralympic Games	CMP2_JOGOS_PARAOLIMPICOS_INVERNO	1
Aquathlon	ESP_AQUATLO	1
Bobsleigh	ESP_BOBSLED	1
Luge	ESP_LUGE	1
Snowboarding	ESP_SNOWBOARD	1
Softball	ESP_SOFTBOL	1

Table 2: Description of *BEST sports* datasets

Dataset	# Docs	# Classes	Class S. D.	Majority Class	Average Silhouette Width	Semantic Complexity
BS-full	881	66	2.05%	10.33%	0.0822	topic
BS-topic	283	4	4.91%	32.16%	0.1978	topic
BS-semantic	283	4	9.04%	32.86%	0.0083	semantic
BS-topic-semantic	283	15	3.85%	16.96%	0.0173	semantic

3. Analysis of *BEST sports* datasets

The work of Sinoara et al. (2017) defined two levels of semantic complexity for document organization problems. The first level (topic organization) consists of the problem of document organization that depends basically on the vocabulary. In this problem, each expected group of documents has its own common terms, so documents can be differentiated mainly by the vocabulary. The second level of semantic complexity (semantic organization) consists of the organization problem that cannot be solved by the vocabulary. It requires a deeper semantic knowledge, as the documents cannot be differentiated.

BS-full and *BS-topic* datasets correspond to topic organization problems, the first level of semantic complexity. It is expected that the documents are organized by sports and some sports have particular terms. For example, among the frequent words in Formula 1 documents are *grid*, *pole*, and *GP*, whereas *set*, *game*, and *match* are among the frequent words in Tennis documents. *BS-semantic* and *BS-topic-semantic* datasets correspond to semantic organization problems, the second level of semantic complexity. Vocabulary is not enough to differentiate a document that reports a victory of a Brazilian athlete from a document that reports a Brazilian defeat.

In order to measure the difficulty of each dataset, Sinoara et al. (2017) analyzed the compactness and separation of the known clusters using the silhouette width criterion. Figures 1 to 4 illustrate the silhouette width of each dataset. Small silhouette width (around 0) indicates that the document lies at a cluster border, near other clusters. We can note that the topic organization datasets (Figures 1 and 2) have average silhouette widths higher than the semantic organization datasets (Figures 3 and 4). Sinoara et al. (2017) discuss this subject.

Figure 1: *BS-full* - Average silhouette width: 0.08 (Sinoara et al., 2017)

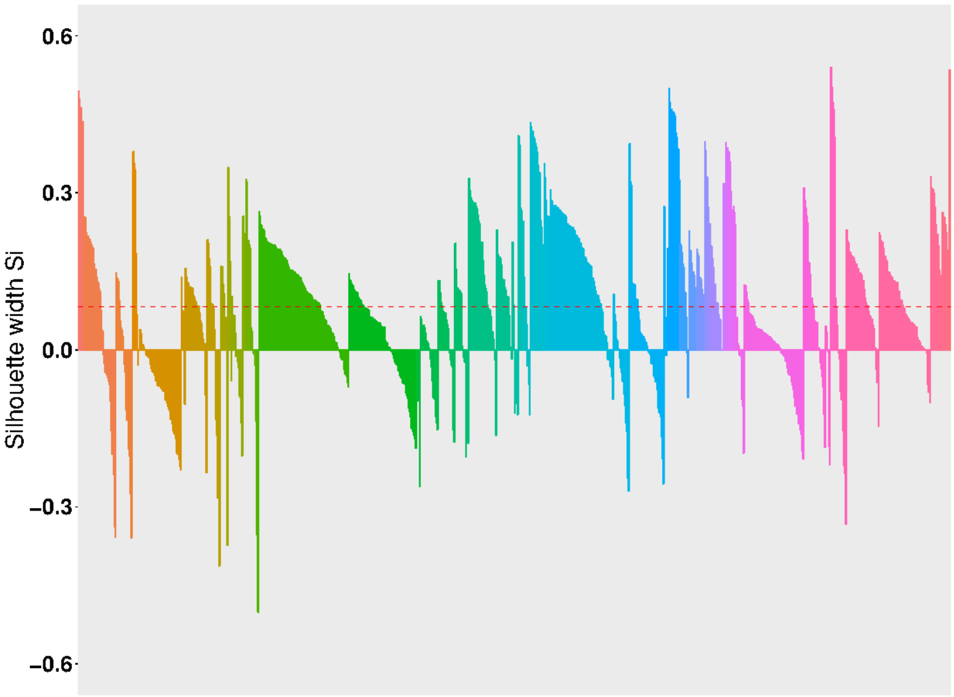


Figure 2: *BS-topic* - Average silhouette width: 0.20 (Sinoara et al., 2017)

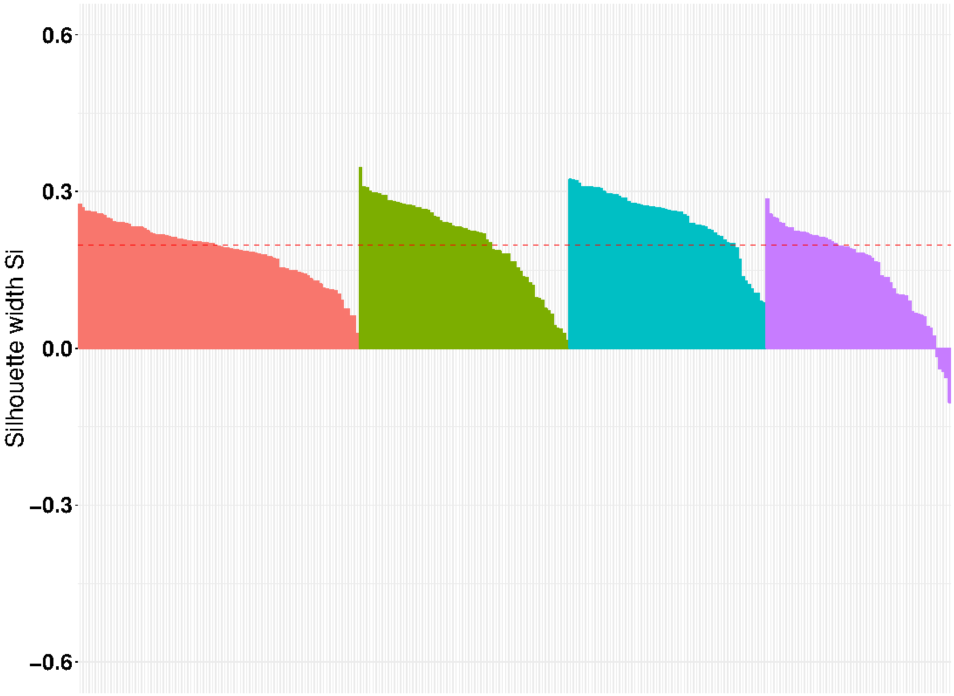


Figure 3: *BS-semantic* - Average silhouette width: 0.01 (Sinoara et al., 2017)

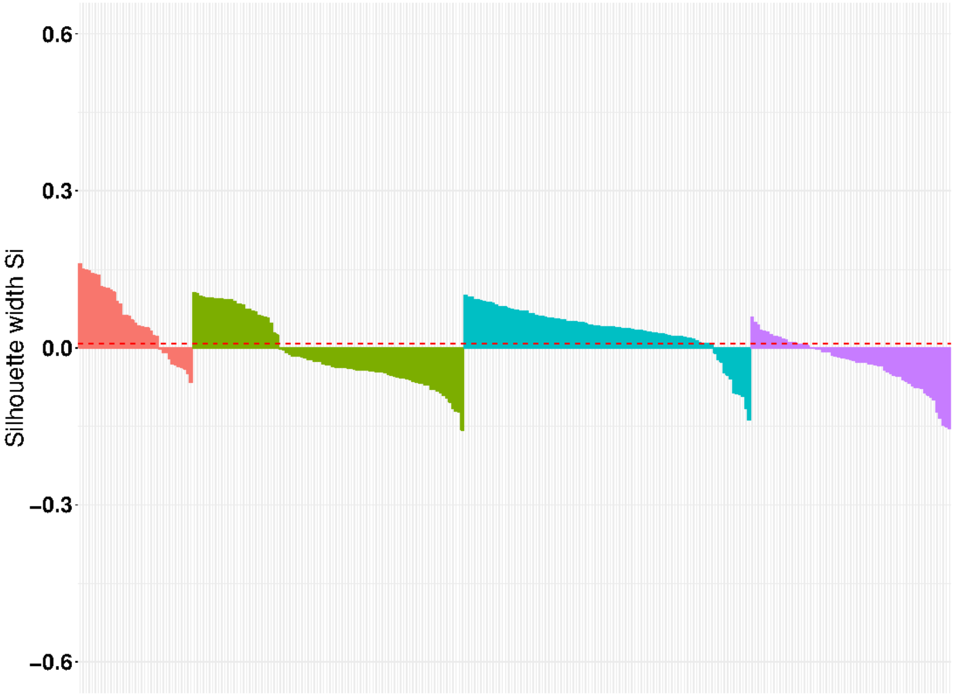
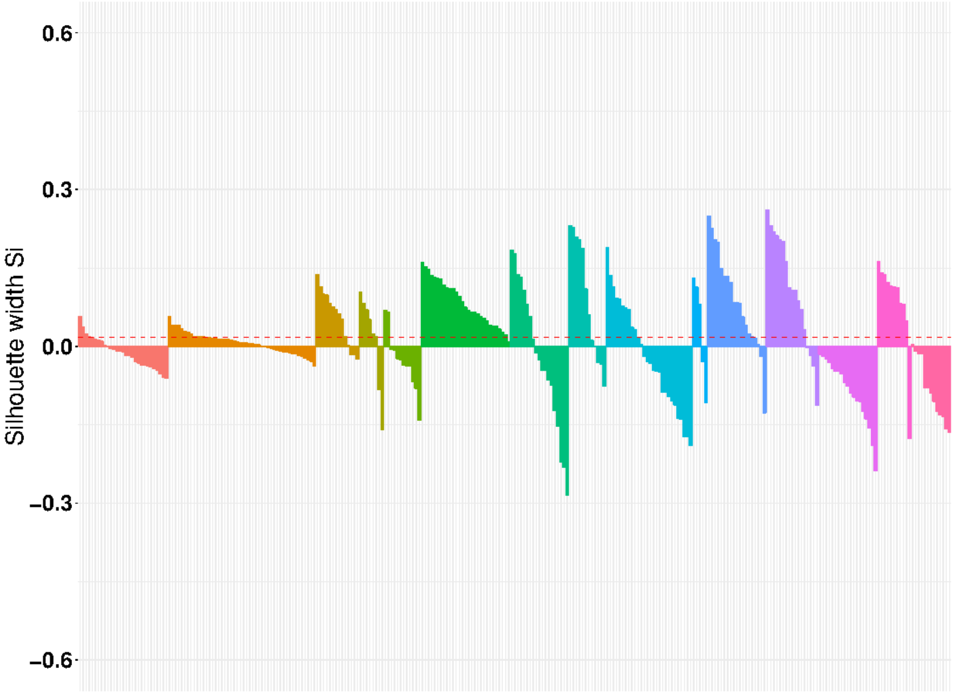


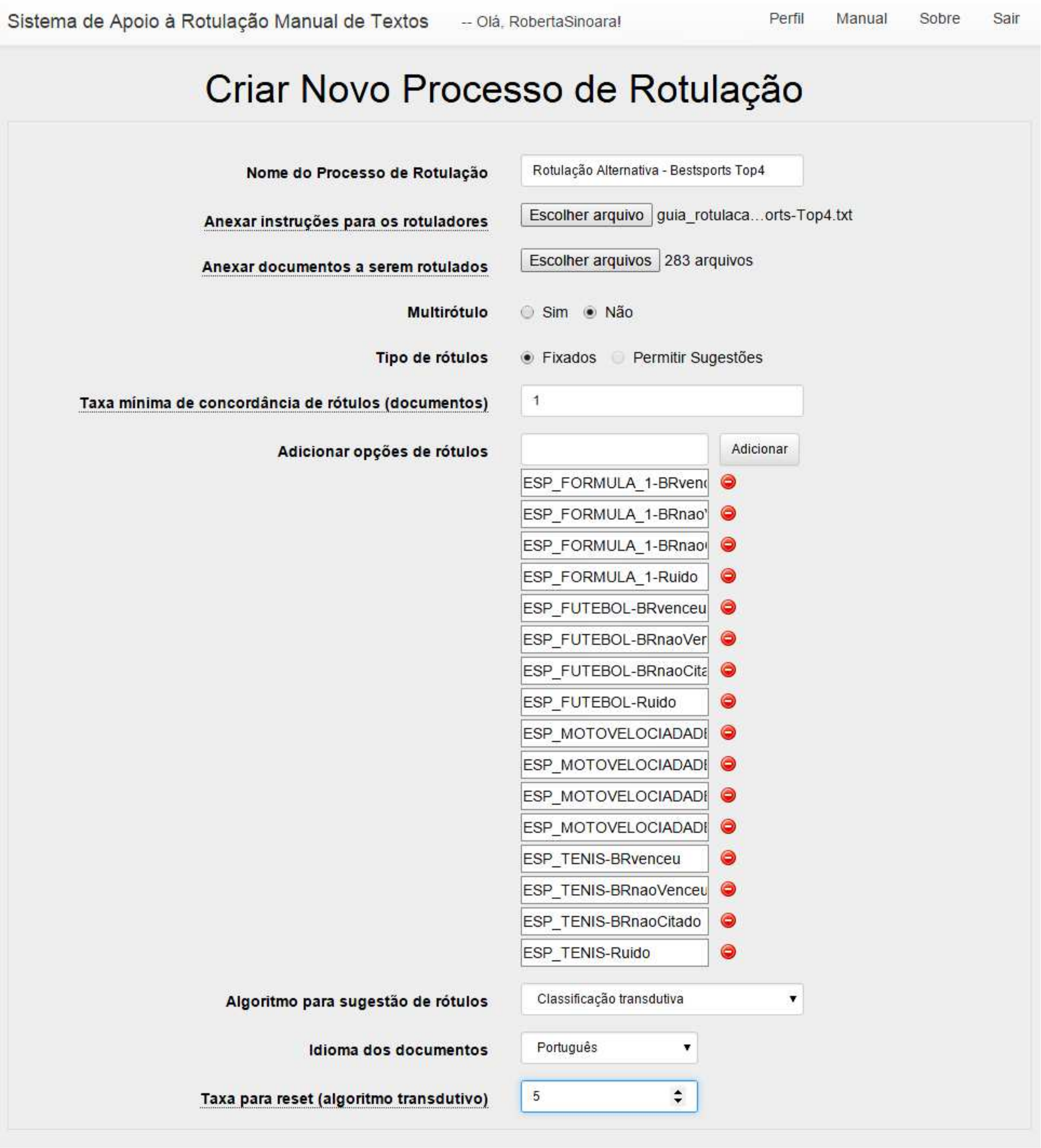
Figure 4: *BS-topic-semantic* - Average silhouette width: 0.02 (Sinoara et al., 2017)



4. Labeling process

The process of manual labeling the *BEST sports - Top 4* collection was supported by RotuLABIC[3] tool (Paravia et al., 2015). Figure 5 presents the configuration parameters set for the RotuLABIC labeling process. The process execution followed a labeling guide (written in Portuguese), which is presented in Figure 6. The objective of this process was to label *BEST sports - Top 4* documents according to the reported performance of Brazilian athletes.

Figure 5: RotuLABIC - Labeling process configuration



Criar Novo Processo de Rotulação

Nome do Processo de Rotulação: Rotulação Alternativa - Bestsports Top4

Anexar instruções para os rotuladores: guia_rotulaca...orts-Top4.txt

Anexar documentos a serem rotulados: 283 arquivos

Multirótulo: ○ Sim ● Não

Tipo de rótulos: ● Fixados ○ Permitir Sugestões

Taxa mínima de concordância de rótulos (documentos): 1

Adicionar opções de rótulos:

- ESP_FORMULA_1-BRvenc
- ESP_FORMULA_1-BRnao'
- ESP_FORMULA_1-BRnao
- ESP_FORMULA_1-Ruido
- ESP_FUTEBOL-BRvenceu
- ESP_FUTEBOL-BRnaoVen
- ESP_FUTEBOL-BRnaoCita
- ESP_FUTEBOL-Ruido
- ESP_MOTOVELOCIADADE
- ESP_MOTOVELOCIADADE
- ESP_MOTOVELOCIADADE
- ESP_MOTOVELOCIADADE
- ESP_TENIS-BRvenceu
- ESP_TENIS-BRnaoVenceu
- ESP_TENIS-BRnaoCitado
- ESP_TENIS-Ruido

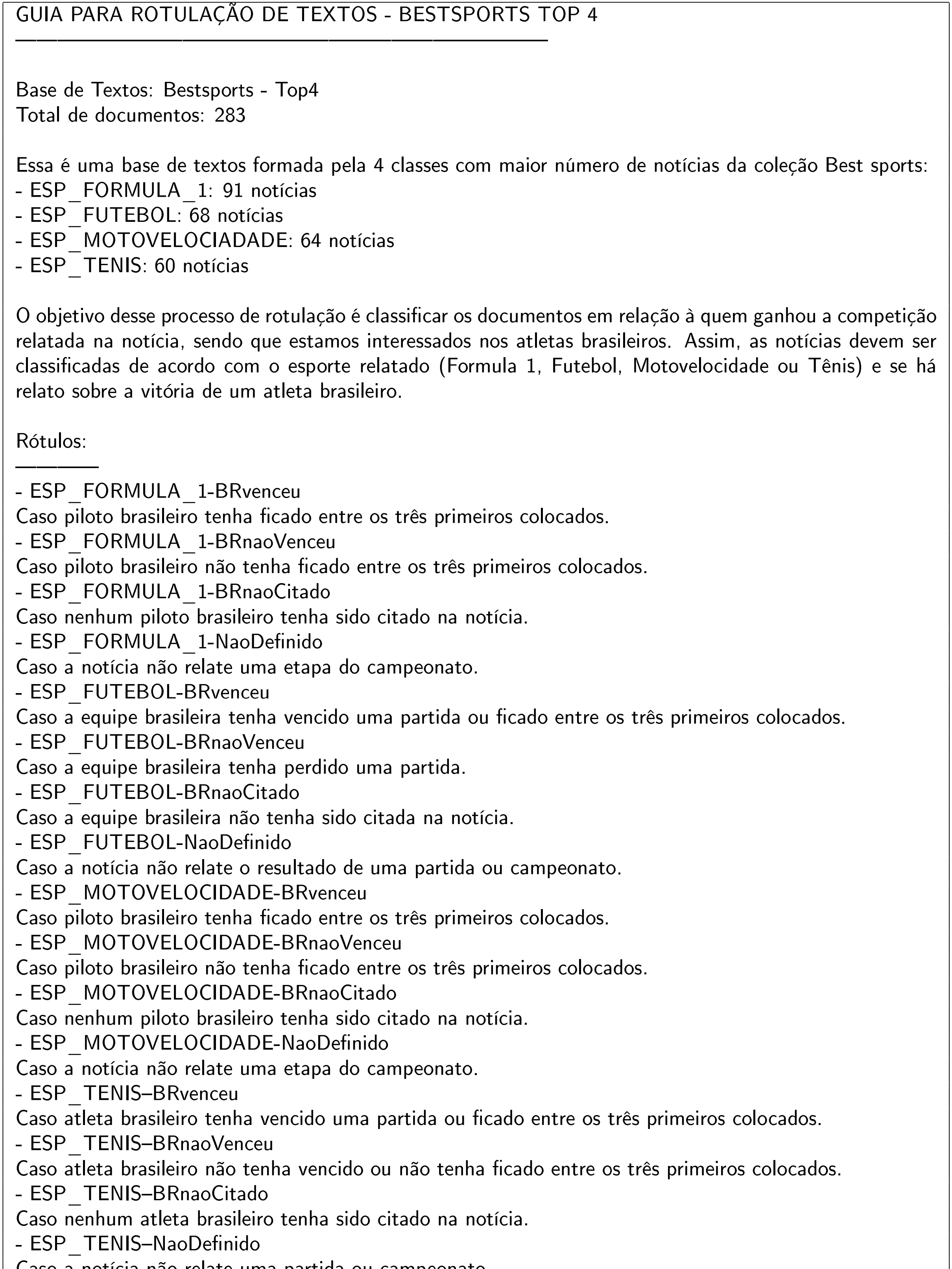
Algoritmo para sugestão de rótulos: Classificação transdutiva

Idioma dos documentos: Português

Taxa para reset (algoritmo transdutivo): 5

[3] RotuLABIC: http://labic.icmc.usp.br/material/14

Figure 6: *BEST sports - Top 4* - Labeling guide (in Portuguese)



GUIA PARA ROTULAÇÃO DE TEXTOS - BESTSPORTS TOP 4

——————————————————————————

Base de Textos: Bestsports - Top4
Total de documentos: 283

Essa é uma base de textos formada pela 4 classes com maior número de notícias da coleção Best sports:
- ESP_FORMULA_1: 91 notícias
- ESP_FUTEBOL: 68 notícias
- ESP_MOTOVELOCIADADE: 64 notícias
- ESP_TENIS: 60 notícias

O objetivo desse processo de rotulação é classificar os documentos em relação à quem ganhou a competição relatada na notícia, sendo que estamos interessados nos atletas brasileiros. Assim, as notícias devem ser classificadas de acordo com o esporte relatado (Formula 1, Futebol, Motovelocidade ou Tênis) e se há relato sobre a vitória de um atleta brasileiro.

Rótulos:

————

- ESP_FORMULA_1-BRvenceu
Caso piloto brasileiro tenha ficado entre os três primeiros colocados.
- ESP_FORMULA_1-BRnaoVenceu
Caso piloto brasileiro não tenha ficado entre os três primeiros colocados.
- ESP_FORMULA_1-BRnaoCitado
Caso nenhum piloto brasileiro tenha sido citado na notícia.
- ESP_FORMULA_1-NaoDefinido
Caso a notícia não relate uma etapa do campeonato.
- ESP_FUTEBOL-BRvenceu
Caso a equipe brasileira tenha vencido uma partida ou ficado entre os três primeiros colocados.
- ESP_FUTEBOL-BRnaoVenceu
Caso a equipe brasileira tenha perdido uma partida.
- ESP_FUTEBOL-BRnaoCitado
Caso a equipe brasileira não tenha sido citada na notícia.
- ESP_FUTEBOL-NaoDefinido
Caso a notícia não relate o resultado de uma partida ou campeonato.
- ESP_MOTOVELOCIDADE-BRvenceu
Caso piloto brasileiro tenha ficado entre os três primeiros colocados.
- ESP_MOTOVELOCIDADE-BRnaoVenceu
Caso piloto brasileiro não tenha ficado entre os três primeiros colocados.
- ESP_MOTOVELOCIDADE-BRnaoCitado
Caso nenhum piloto brasileiro tenha sido citado na notícia.
- ESP_MOTOVELOCIDADE-NaoDefinido
Caso a notícia não relate uma etapa do campeonato.
- ESP_TENIS–BRvenceu
Caso atleta brasileiro tenha vencido uma partida ou ficado entre os três primeiros colocados.
- ESP_TENIS–BRnaoVenceu
Caso atleta brasileiro não tenha vencido ou não tenha ficado entre os três primeiros colocados.
- ESP_TENIS–BRnaoCitado
Caso nenhum atleta brasileiro tenha sido citado na notícia.
- ESP_TENIS–NaoDefinido
Caso a notícia não relate uma partida ou campeonato.

Table 3 presents some specific details of each sports category and Table 4 presents the class distribution of *BEST sports - Top 4* collection, according to the classification of *BS-topic-semantic* dataset. It must be noted that the class label NaoDefinido (*"Not defined"*) was initially labeled as Ruido (*"Noise"*). After the labeling process, it was renamed to improve its label meaning.

Table 3: Specific details of each sports category

Sports category	Notes
Fórmula 1 (Formula 1)	- Any podium position is considered a victory. - Qualifying session results (grid position) are treated as competition results. - Documents reporting season information and car launches are labeled as ESP_FORMULA_1-NaoDefinido. - Documents reporting pre-season testings are labeled as ESP_FORMULA_1-NaoDefinido.
Futebol (Soccer)	- The focus is results of matches. - Tournament rankings are not considered victory or defeat. - Documents reporting ties of Brazilian team matches are labeled as ESP_FUTEBOL-NaoDefinido. - Documents reporting FIFA *Ranking* are labeled as ESP_FUTEBOL-NaoDefinido.
Motovelocidade (MotoGP)	- Formula 1 notes also apply to MotoGP. - Documents reporting provisional pole position are labeled as ESP_MOTOVELOCIDADE-NaoDefinido.
Tênis (Tennis)	- The focus is results of matches. - Documents reporting both victory and defeat of a Brazilian athlete are labeled as ESP_TENIS–NaoDefinido.

Table 4: Class distribution of *BEST sports - Top 4* collection

Class label	# Docs
ESP_FORMULA_1-BRvenceu	48
ESP_FORMULA_1-BRnaoVenceu	29
ESP_FORMULA_1-BRnaoCitado	0
ESP_FORMULA_1-NaoDefinido	14
ESP_FUTEBOL-BRvenceu	29
ESP_FUTEBOL-BRnaoVenceu	12
ESP_FUTEBOL-BRnaoCitado	8
ESP_FUTEBOL-NaoDefinido	19
ESP_MOTOVELOCIDADE-BRvenceu	5
ESP_MOTOVELOCIDADE-BRnaoVenceu	28
ESP_MOTOVELOCIDADE-BRnaoCitado	12
ESP_MOTOVELOCIDADE-NaoDefinido	19
ESP_TENIS–BRvenceu	11
ESP_TENIS–BRnaoVenceu	19
ESP_TENIS–BRnaoCitado	17
ESP_TENIS–NaoDefinido	13

5. Final remarks

In this technical report, we described the *BEST sports* collection and the four derived datasets. The document collection was collected and prepared to be used as benchmarking collection in text mining research. We created the datasets referring to problems of different semantic complexity levels. *BEST sports* documents are available at `http://sites.labic.icmc.usp.br/rsinoara/bestsports`. The *BEST sports* collection is provided for non-commercial and research purposes only. If you make use of this collection or any derivative of it, please, consider citing this technical report.

Acknowledgements

We would like to thank Nabil Jasser Husein, owner and editor of BEST sports website, who kindly provided his news articles to build the *BEST sports* collection for research purposes.

This work was supported by FAPESP: grant #2013/14757-6, São Paulo Research Foundation (FAPESP). The opinions, hypotheses and conclusions or recommendations expressed in this material are the authors' own and do not necessarily reflect the views of FAPESP.

References

Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications* (1st ed.). Chapman & Hall/CRC.

Aggarwal, C. C. & C. Zhai (Eds.) (2012). *Mining Text Data*. Springer.

Paravia, R. d. P. P., R. A. Sinoara, R. G. Rossi, & S. O. Rezende (2015). Sistema para apoio à rotulação manual de textos utilizando aprendizado de máquina. In *SIICUSP: Anais do 23o. Simpósio Internacional de Iniciação Científica da USP*.

Sinoara, R. A., R. B. Scheicher, & S. O. Rezende (2017). Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity. In *CIDM'17: Proceedings of the 2017 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 2057–2064.