Exploring multilingual news articles by combining Sentiment Analysis classifiers

Caio Mello - School of Advanced Study, University of London, London, United Kingdom | caio.mello@sas.ac.uk
Gullal S. Cheema - TIB Leibniz Information Center for Science and Technology, Hannover, Germany | gullal.cheema@tib.eu
Gaurish Thakkar - Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia | gthakkar@m.ffzg.hr

Abstract This study aims to present an approach for the challenges of working with Sentiment Analysis (SA) applied to news articles in a multilingual corpus. It looks at the use and combination of multiple algorithms to explore news articles published in English and Portuguese. As a case study, the method was applied to the study of the media coverage of London 2012 and Rio 2016 Olympic legacies.

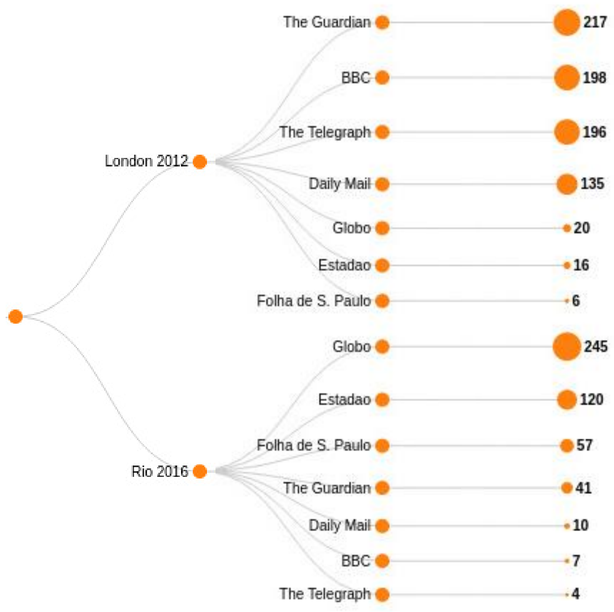
Introduction Sentiment Analysis (SA) is a Natural Language Processing (NLP) technique to identify emotions discursively expressed in a text (Liu, 2012). This work aims to apply Sentiment Analysis algorithms to study news articles. This journalistic text genre presents specific challenges due to the form and content such as the length of the text and the minimised use of adjectives. This requires some effort from researchers to adapt traditional approaches produced for the use in subjective statements like Twitter (Shirsat et al., 2017; Balahur, 2012; Taj, 2019).

Case Study: the Olympic legacy of London 2012 and Rio 2016 Although highly associated with a positive meaning, the word legacy describes not only the gains of the cities by hosting the event but all the aspects that emerge from this process, whether 'planned or unplanned, positive or negative, intangible or tangible' (Gratton and Preuss, 2008, p. 1924).

Six years later, the wonderful legacy of London 2012 (Globo, 2018)

Rio's Olympic legacy a 'huge disappointment' (BBC News, 2017)

Data Overview



- 1271 news articles
- seven media outlets
- Portuguese and English
- London 2012 and Rio 2016

Evaluating the classifiers

	PT Sentic Net	PT Senti Strength	PT Vader	PT Twitter BERT	EN Sentic Net	EN Senti Strength	EN Vader	EN Amazon BERT	EN Sent140 BERT
Rio_Globo	33%	39.5%	49.7%	24.5%	46.5%	50.2%	55.5%	55.5%	61.2%
Rio_Estadao	26.4%	15.7%	39.6%	3.3%	38.8%	31.4%	40.4%	53.7%	57%
London_Guardian	x	x	x	x	48%	49.6%	56.2%	60%	58.5%
London_DailyMail	x	x	x	x	52%	47.9%	56.6%	61.7%	61.2%

Table 1 - (%) Matches of each classifier with the gold labels (accuracy evaluation for this dataset) for news headline text. PT: Portuguese; EN: English. Total of 717 news headlines (56.4% of the total number of articles).

Combining classifiers

(A) Combination of five sentiment classifiers (majority agreement)

(B) Combination of top three sentiment classifiers (discarding two worst scores)

(C) Combination of three classifiers replacing inconclusive by Sent140 BERT

(D) Combination of three classifiers ignoring inconclusive sentences

Method	Rio_Globo	Rio_Estadao	London_Guardian	London_DailyMail
Sent140 BERT	61%	57%	61%	58.5%
Approach A	59%	48.3%	59.4%	53.3%
Approach B	62%	57.5%	66%	57.7%
Approach C	65.3%	59.16%	70%	61.4%
Approach D	**66.9%**	**65%**	**74.7%**	**69.6%**
Inconclusive	7.3%	11.6%	10.5%	17%

Table 2 - (%) Accuracy of each approach for news headline text.

Potentialities and limitations of the method SHAP (Lundberg & Lee, 2017) used Shapley values to depict how fairly the outcome can be distributed among the different features.

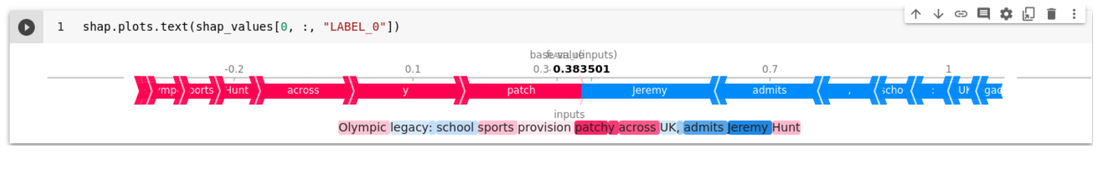


Figure 1 - Shap returns a pictorial depiction of Shapley values. The negative sentiments are represented in red. The positive ones, in blue. The label assigned to this sentence by the combination of classifiers was positive.

Category	News headline
Entity	Olympic legacy: school sports provision patchy across **UK**, admits **Jeremy Hunt**
Punctuation	**The Essential Morning:** the melancholy inheritance of the Olympics
Semantic	Environmental legacy, the **great** debt of Rio Olympics
Negation	Olympic stadium will **not** be white elephant after London 2012
Metaphor	Letters: The true Olympic legacy is **white elephants** on our doorstep
Domain specific	Britain's Olympic legacy is a **sedentary** nation | David Conn
Sarcasm	Martin Samuel: Tessa Jowell deserves an **Olympic medal in utter madness**
Translation	Condominiums **spare** (DUMP) sewer irregularly at Jacarepaguá **Lagoon** (LAKE)
Conjunctions	West Ham's Olympic Stadium move means we still have a home for heroes **but** it doesn't sit there doing nothing for 50 weeks
Objective statements	London 2012 Olympics will **cost a total of £8.921bn**, says minister

Table 3 - (%) Accuracy of each approach for news headline text.

Impacts of Translation

Portuguese (original)	English (translation)
Brasil anuncia ao COI **corte** de R$ 900 mi no orçamento dos Jogos	Brazil announces the COI **Court** of R$ 900 mi in the games budget
Legado olímpico virou milhões em propina a 'amigos da **corte**' de Cabral, diz MPF	Olympic legacy turned millions in tipping '**Cut** Friends' of Cabral, says MPF
Olimpíada é 'desculpa fantástica' para mudar o **Rio**, diz prefeito	Olympiad is 'fantastic excuse' to change the **river**, says mayor

Table 4 - Comparison between the original sentences in Portuguese and their translations into English.

Interpreting sentiments

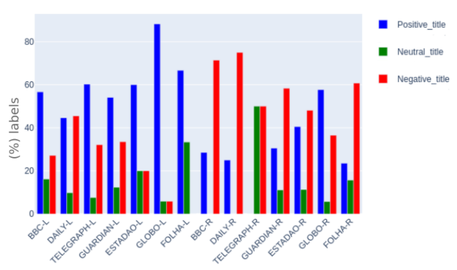


Chart 1 - (%) Comparison between the percentage of positive, negative and neutral labels assigned to news titles. (L) London; (R) Rio.

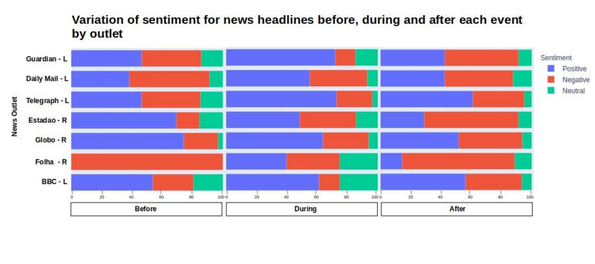


Chart 2 - (%) Variation of sentiment for news headlines before, during and after each event by outlet. Timeframe for London 2012 Olympics: before (2004 to 2011), during (2012), after (2012 to 2020). Timeframe for Rio 2016 Olympics: before (2009 to 2015), during (2016) and after (2017 to 2020).

Conclusion The experiments have shown that applying Vader, Amazon BERT and Sent140 BERT to the text corpus and extracting the agreements between at least two classifiers as a correct answer has considerably increased the accuracy of SA results for this specific data. Although mistakes of translation had impacted the final sentiment classification, the overall outputs were considerably better than the use of any of the listed modules for Portuguese. By comparing the coverage of the two events, we concluded that London received a more positive coverage while Rio a more negative one. While the Brazilian media has been less critical about London, the sentiments expressed by the British media about Rio were very negative. We have referred to this phenomenon as a utopian-dystopian dichotomy, where one event is represented as a 'huge disappointment' while the other one as a 'success', silencing about or reducing the space in the agenda for the nuances embedded in complex media events like these ones.

References

Balahur, A., & Turchi, M. (2012). Multilingual sentiment analysis using machine translation. In Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis (pp. 52-60).

Gratton, C., & Preuss, H. (2008). Maximizing Olympic Impacts by Building Up Legacies. The International Journal of the History of Sport, 25(14), 1922–1938. https://doi.org/10.1080/09523360802439023

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30. Available on https://github.com/slundberg/shap. Accessed on 10th January, 2022.

Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017). Document Level Sentiment Analysis from News Articles. 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 1–4. https://doi.org/10.1109/ICCUBEA.2017.8463638

Taj, S., Shaikh, B. B., & Meghji, A. F. (2019). Sentiment analysis of news articles: a lexicon based approach. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-5). IEEE.

CLEOPATRA
cleopatra-project.eu
Cross-lingual Event-centric Open Analytics Research Academy
Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997.
SCHOOL OF ADVANCED STUDY | UNIVERSITY OF LONDON