MARIANA GONÇALVES DA COSTA[1]
Universidade Federal do Rio de Janeiro, PPGI, Rio de Janeiro, Brazil

MATHEUS DO Ó SANTOS TIBURCIO[2]
Universidade Federal do Rio de Janeiro, PPGI, Rio de Janeiro, Brazil

MARCIA DOS SANTOS MACHADO VIEIRA[3]
Universidade Federal do Rio de Janeiro, R/PPGLEV, io de Janeiro, Brazil

SÉRGIO MANUEL SERRA DA CRUZ[4]
Universidade Federal do Rio de Janeiro, PPGI, Rio de Janeiro, Brazil

Challenges in Multiword Expressions Processing for NLP Tasks in Portuguese

ABSTRACT
In Natural Language Processing, Multiword Expressions (MWEs) refer to combinations of words with idiosyncratic properties; that is, their meaning cannot be directly inferred by the meaning of the individual words and are usually interpreted as a single lexical unit. Among MWEs, some expressions can function idiomatically and literally, requiring comprehension of both meanings. For example, the expressions *morrer de/a X* (to die of X) and *chorar de/a X* (to cry of X) in Portuguese are commonly used metaphorically for intensification, but can also be interpreted literally. This semantic ambiguity poses challenges for NLP tasks, such as sentiment analysis, that depend on the reliable interpretation of figurative language. These challenges are especially pronounced in low-resource languages, where the scarcity of annotated data is another issue. This paper presents an ongoing study investigating the impact of MWEs on sentence-level sentiment analysis in Portuguese, with a particular focus on openly available BERT-based models. Beyond evaluating model performance, we aim to demonstrate how insights from theoretical linguistics can inform NLP research. Relying on Usage-Based Construction Grammar, we analyze MWEs as instances of constructions, approaching them as conventionalized meaningful patterns rather than isolated lexical units. Preliminary results show that metaphorical intensifier constructions significantly interfere with sentiment classification in the trained models available for Portuguese. These findings highlight the importance of incorporating linguistically informed representations of meaning variation and constructional patterns into NLP methodologies.

KEYWORDS
Multiword Expressions, Sentiment Analysis, Natural Language Processing, Digital Humanities

---

[1] https://orcid.org/0000-0002-8088-0794, marianag.costta@gmail.com, CAPES/DS
[2] https://orcid.org/0009-0004-4985-8291, mathh2899@gmail.com
[3] https://orcid.org/0000-0002-2320-5055 , marcia@letras.ufrj.br, FAPERJ/CNE and CNPq Researcher
[4] https://orcid.org/0000-0002-0792-8157, sergioserra@ic.ufrj.br

1. Introduction

Natural Language Processing (NLP) is a field that studies the development of computational methods for processing human language. NLP-based tools such as machine translation systems, search engines, chatbots, spam filters, and spell checkers are increasingly embedded in everyday life, often operating invisibly as digital technologies become more integrated into daily activities. Although NLP is interdisciplinary by definition, combining knowledge from linguistics, computer science, and artificial intelligence (Johri *et al.*, 2021), linguists' participation in the area is restricted, often confined to data collection and annotation. The rise of large language models (LLMs), however, renewed interest from linguists in how these systems represent and reproduce human language, while computer scientists became more interested in linguistics. This scenario creates valuable opportunities for collaboration, which are essential for developing NLP systems that better capture the diversity and nuances of human language.

Despite recent advances, many NLP tasks continue to struggle with the underlying meaning of language usage. Trott *et al.* (2020) argue that most approaches to language meaning in NLP are based on objective information; that is, they do not take into account subtle linguistic choices that convey speakers' perspectives, emphasis, and opinions. Neglecting these choices can lead to significant issues, particularly in applications that heavily rely on semantic information, such as sentiment analysis, which aims to determine the emotional tone or sentiment expressed in a text. The challenge of evaluating feelings in the presence of figures of speech, such as irony, metaphor, or even simple negation, has been widely described in the NLP literature (Kiritchenko *et al.*, 2016; Wankhade *et al.*, 2022). The distance between the intended meaning and what is objectively said should not be treated as an exception in language use, since the metaphorical and creative use of language is natural to human communication (Lakoff & Johnson, 1980).

Multiword Expressions (MWEs), for instance, are present in both informal and formal contexts (e.g., "put up with", "a piece of cake", "kick the bucket"). They can be defined as combinations of words with idiosyncratic properties; that is, their meaning cannot be directly inferred by the meaning of the individual words that compose them and are usually interpreted as a single lexical unit. MWEs pose a significant challenge for NLP analysis, as they require an approach to text segmentation and semantic analysis that go beyond word-by-word processing. Moreover, their composition tends to be highly metaphorical, and the linguistic choices involved may convey meanings that diverge sharply from – or even contradict – the objective information present in the text.

One prominent example is the use of verbs related to negative events to intensify feelings or situations. In Portuguese, the verbs *chorar* ("to cry") and *morrer* ("to die") are commonly used in expressions of intensification, as noted in the literature (Mota, 2021; Silva, 2024). When followed by the preposition "*de*" or "*a"* (in European Portuguese), these verbs can act as intensifiers of an event or feeling in Portuguese. For instance, one might say:

(a) Eu **to chorando de rir** das imitações que @USER faz dos cantores!![5]
I'm crying laughing at the imitations @winderson_nunes does of the singers!!

(b) **Morrendo de amores** pelo @USER no programa do @USER. #TheNoite
I'm dying of love for @USER on @USER's show. #TheNoite

Similarly, these verbs also appear in literal forms, as illustrated below:

[5] More information about the dataset can be found in the Section 2.

(c) Tão lindinho ver eles **chorando de felicidade e orgulho** de si. #MasterChefBR
It's so sweet to see them crying of happiness and pride. #MasterChefBR

(d) Grécia: "Não vamos deixar nenhum dos nossos concidadãos **morrer de frio**", afirmou o ministro do Ambiente, Yannis...
Greece: "We will not let any of our fellow citizens die of cold," stated the Minister of the Environment, Yannis...

The possibility of using these expressions in both literal and figurative senses poses unique challenges for language processing. In sentence-level sentiment analysis, a lexicon-based algorithm is likely to incorrectly assign negative sentiment values to *chorar* and *morrer*, even when they function as intensifiers in the expressions *chorar de/a X* and *morrer de/a X*. Lexicon-based (or rule-based) approaches represent a traditional method of sentiment analysis in which sentiment values are assigned to individual words based on a predefined dictionary (e.g., *good* as positive and *bad* as negative), and then aggregated to produce an overall sentiment score for the text. This approach usually evaluates the sentences literally, struggling with figurative uses.

By contrast, machine learning approaches typically rely on supervised learning, in which a classifier is trained on previously annotated data to determine the polarity of new texts, and, therefore, are better equipped to deal with metaphorical language. Although machine learning methods generally achieve higher accuracy in sentiment analysis, they require large amounts of training data and programming expertise, and their outputs are often less transparent or interpretable (Taboada, 2016). With the rise of large language models, however, the application of machine learning techniques in NLP has become considerably more accessible. Pretrained models are now openly available online and can be applied with minimal programming knowledge, making sentiment analysis and other NLP tasks more accessible.

The widespread adoption of Transformer-based models in NLP is largely attributed to their ability to capture local and global contextual information during text processing and generation. Unlike earlier models, which process words sequentially, Transformers analyze entire sequences of words simultaneously, which enables effective modeling of long-distance dependencies and complex relationships between words. This enhanced context awareness may allow such models to more accurately classify instances of the MWEs discussed above, as their interpretation depends on the immediate context. BERT (*Bidirectional Encoder Representations from Transformer*s, Devlin *et al.*, 2019) is a widely used model in NLP that can be trained locally and easily replicated, making it particularly suitable for controlled experimental settings, thereby supporting reproducibility in research. To test this hypothesis, the present study is structured as follows:

(i) the selection of Portuguese-language sentiment analysis datasets;
(ii) the application of two openly available BERT models trained on sentiment analysis of Portuguese; and
(iii) a qualitative and quantitative evaluation of these models.

Along these lines, this study aims to foster collaboration between linguistics and computer science, focusing on the interdisciplinary quality of Natural Language Processing. While computational methods provide scalability, efficiency, and formal modeling capabilities, linguistic theory contributes with explanatory depth, empirical grounding, and principled descriptions of language phenomena. This collaboration is particularly crucial in the processing of Multiword Expressions, which challenge traditional assumptions of compositionality and word-based processing. In this context, adopting Construction Grammar

as a theoretical framework offers significant advantages for the description and analysis of MWEs in Portuguese. Construction Grammar conceptualizes language as a network of form–meaning pairings at multiple levels of abstraction, including fixed, semi-fixed, and schematic expressions, without forcing a strict dichotomy between lexicon and grammar. This perspective is especially well suited to Portuguese MWEs, many of which display gradient idiomaticity, syntactic variability, and pragmatic constraints that cannot be adequately captured by purely lexical or rule-based models. By aligning linguistic insights from Construction Grammar with computational representations, this interdisciplinary approach enables more linguistically informed NLP systems, fostering analyses that are both theoretically motivated and computationally effective.

2. Dataset

The data selection process was largely guided by the study *Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark* by Augustyniak *et al.* (2023). In this work, the authors present the MMS Dataset and Benchmark, currently the largest open-access dataset for sentiment classification. Most importantly, they describe in detail the criteria used to select datasets to be included in the MMS collection, including: (i) robust annotation; (ii) well-defined annotation guidelines; (iii) numeric classification; (iv) annotation in three classes, excluding binary labeling; and (v) monolingual datasets, dividing multilingual corpora in monolingual collections when necessary.

From this collection, we were able to extract 157,393 tweets from European Portuguese collected between October and December 2013 (Mozetič *et al.*, 2016) and 15,000 from Brazilian Portuguese (Brum & Volpe Nunes, 2018), both part of the MMS Dataset and Benchmark. The first dataset was originally part of a multilingual *corpus* for Twitter sentiment classification, and all Portuguese tweets were labeled by a single annotator with a performance of *Krippendorff's Alpha* ≈ 0.391. The second dataset is TweetSentBr, a widely-adopted sentiment classification *corpus* of tweets in Brazilian Portuguese on TV show domain. The data was extracted from Twitter between January and July 2017, and labeled by seven annotators following guidelines with examples, definitions and tips for the annotation process.

Using a web-based dataset is particularly beneficial to this study due to its informal characteristics and the presence of direct interactions between speakers, containing examples of language used in natural conversions, including slangs, idiomatic expressions, and metaphors. It is worth mentioning, however, that Portuguese remains considerably underrepresented in the MMS Dataset and Benchmark. The lack of data from the domains of news and reviews directly impacts the evaluation of sentiment classification models for Portuguese. This limitation emphasizes both the relevance of this study and the urgency of collaborative efforts to remove Portuguese from the group of low-resource languages.

2.1 Dataset Cleaning

By applying the regular expression *\b(morr|chor)\w*\s+(de|a)\b*, we extracted instances of the intensifier constructions *morrer de/a X* and *chorar de/a X* from the dataset, resulting in 331 rows of data. European and Brazilian Portuguese were not aggregated, in order to allow for a comparison between the two linguistic varieties. Following the extraction process, we examined the labeling and identified a considerable number of inconsistencies, despite the guidelines followed by both the annotation team and the MMS curation team. A few examples of incorrect labeling are:

(e) Começar o dia com a @ANAMARIABRAGA vestida de unicórnio é mesmo surreal. **Morrendo de rir**! #MAISVOCÊ
Starting the day with @ANAMARIABRAGA dressed as a unicorn is truly surreal. Dying from laughter! #MAISVOCÊ
Original label: neutral **Correct label**: positive **Source**: PTBR

The following example shows a misunderstanding of the task. In sentiment analysis, the classification focuses on identifying the evaluation or opinion expressed about an event, object, or agent. Therefore, factual information is not considered negatively unless the speaker explicitly conveys a negative attitude or judgment.

(f) Em 2010, mais de 220 mil pessoas **morreram de cancro de pulmão** provocado pela poluição do ar. #poluição
In 2010, more than 220,000 people died from lung cancer caused by air pollution. #pollution
Original label: negative **Correct label**: neutral **Source**: PTEU

(g) É possível **morrer de tédio**?
Is it possible to die from boredom?
Original label: positive **Correct label**: negative **Source**: PTEU

After identifying these inconsistencies, we manually changed the labels prior to the experiment. All changes were thoroughly annotated, and are available at our GitHub page along with the original data, the experiment results, and all the code used in this study[6].

3. Experimental Design

The experiments involved the selection of two open-access BERT-based models specifically trained for sentiment analysis in Portuguese. Our goal was to evaluate models that were easily accessible to the public, gathering information on the performance of these models as a baseline for future experiments. The first model, mmBERT small multilingual sentiment[7], comprises 100M parameters and was trained using supervised learning on sentiment-annotated data drawn from social media sources, including *tweets*. The similarities of domains represent an important aspect in machine-learning models. The second model, DistilBERT base multilingual cased sentiments[8], contains 135M parameters and was trained via transfer learning from the teacher model *mDeBERTa-v3-base-mnli-xnli* (which relies on multilingual inference data translated from English; however, the trained data from the DistilBERT model is not specified).

Both models were applied to the same manually annotated Portuguese dataset in order to ensure comparability. Their performance was evaluated using standard classification metrics, allowing for a systematic comparison of their effectiveness in classifying sentiment in Portuguese and in handling the linguistic phenomena investigated.

4. Experimental Results

4.1 The usage of morrer and chorar as intensifiers

[6] Available at: <https://github.com/PLaNares-UFRJ/sent-analysis-of-mwes> Access: 28 jan 2025
[7] Available at: <https://huggingface.co/clapAI/mmBERT-small-multilingual-sentiment> Access: 22 jan 2025
[8] Available at: <https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student> Access: 22 jan 2025

In our dataset, the construction with *morrer* was 6 times more productive than that with *chorar*, and the verb *rir* ("to laugh") was the most frequently intensified word. The *chorar* construction, however, exhibits a higher degree of semantic ambiguity, as even when the slot is filled by a positive noun, the expression may still be interpreted literally, as in *chorando de felicidade* ("crying of happiness"). It is worth noting that, in this example, the construction is still acting as an intensifier and, therefore, does not carry negative connotations.

Taking this framework into account, the behavior of these intensifier constructions can be understood as strongly dependent on the event being intensified, whether the expression is used idiomatically or literally. This suggests that, when properly designed, language models fine-tuned for sentiment analysis should not be significantly affected by the idiomatical or literal use of these constructions. Specifically, when applying sentiment analysis, one must take into consideration the context in which these expressions are used; in other words, language usage must be regarded as more than individual words to which 'positive' or 'negative' values are attributed, but as combinations of constructions of different levels of abstraction that interact with each other.

4.2 Evaluation of the Models

The evaluation of the models reveals several shared limitations and complementary strengths, with the accuracy of 0.57 in both models. Both mmBERT and DistilBERT struggled to accurately classify sentiment in the presence of Multiword Expressions, particularly when these constructions interacted with contextual cues such as negation, which frequently led to misclassification. A comparative analysis shows that mmBERT achieved higher accuracy in identifying positive sentiments, whereas DistilBERT performed better on negative sentiment classification (Verify figures 1 and 2), in which DistilBERT showed a tendency to label examples as negatives.

Regarding the type of mislabelling committed by each model, DistilBERT often incorrectly attributed negative values in the presence of the MWEs (See (h)) but also incorrectly attributed positive values (See (i)), indicating a highly arbitrary result.

(h) Amo O @Louro_Jose Ele é show **morrendo de rir**
I love @Louro_Jose. He is amazing, I'm dying from laughter
Target label: positive **Predicted label:** negative **Source:** PTBR

(i) Eu estou a **morrer de calor**!
I'm dying of heat!
Target label: negative **Predicted label:** positive **Source:** PTEU

Meanwhile, mmBERT correctly labeled most of the data from Brazilian Portuguese (accuracy of 0.82) with a few cases of false negatives due to the presence of the MWEs. However, its performance in European Portuguese was considerably worse (accuracy of 0.54), which highly influenced the overall accuracy as most of our data was from European Portuguese. The mmBERT's performance on negative instances was highly affected by the following use of the *morrer* expression:

(j) Estou a **morrer de sono**
I'm dying of sleepiness
Target label: neutral **Predicted label:** negative **Source:** PTEU

This example highlights the complexity and ambiguity often involved in identifying sentiment. The combination *morrer de sono* was commonly found in our dataset; however, its value depended on its use context. For instance, when "morrer de sono" was associated with a task/obligation, such as in "Tenho que voltar ao estudo mas estou a morrer de sono" (*I have to get back to studying, but I'm dying of sleepiness*), the target label was negative, and when it was not linked to a task/obligation, the target label was neutral. Nevertheless, we consider this expression as highly ambiguous, and it requires further evaluation.

Another interesting point in our results was that both models were not consistent in their evaluation. In our sample, there were instances that were extremely similar to others; however, that did not lead to the same labeling as one would expect. For instance, in some cases the model would evaluate *morrer de rir* (cry of laughter) with a positive label, and in a different case where very little information was different with a negative label. This shows that both models do not recognize the function of the expressions as intensifiers.

Additionally, the quality of the labeled data applied for the evaluation, especially in European Portuguese, was a critical issue, as annotation inconsistencies indicate the need for further review and clearer guidelines. These inconsistencies show the complexity of identifying sentiment in written text, a subject commonly addressed in linguistics. It is worth mentioning that TweetSentBr was included in mmBERT's training (which may explain its high performance in Brazilian Portuguese) and may have been included in DistilBERT, though its training data are not fully disclosed. Beyond dataset-related concerns, both models suffer from limited transparency regarding data provenance and training procedures, which jeopardizes interpretability and systematic error analysis.

FIGURE 1
Confusion Matrix - mmBERT

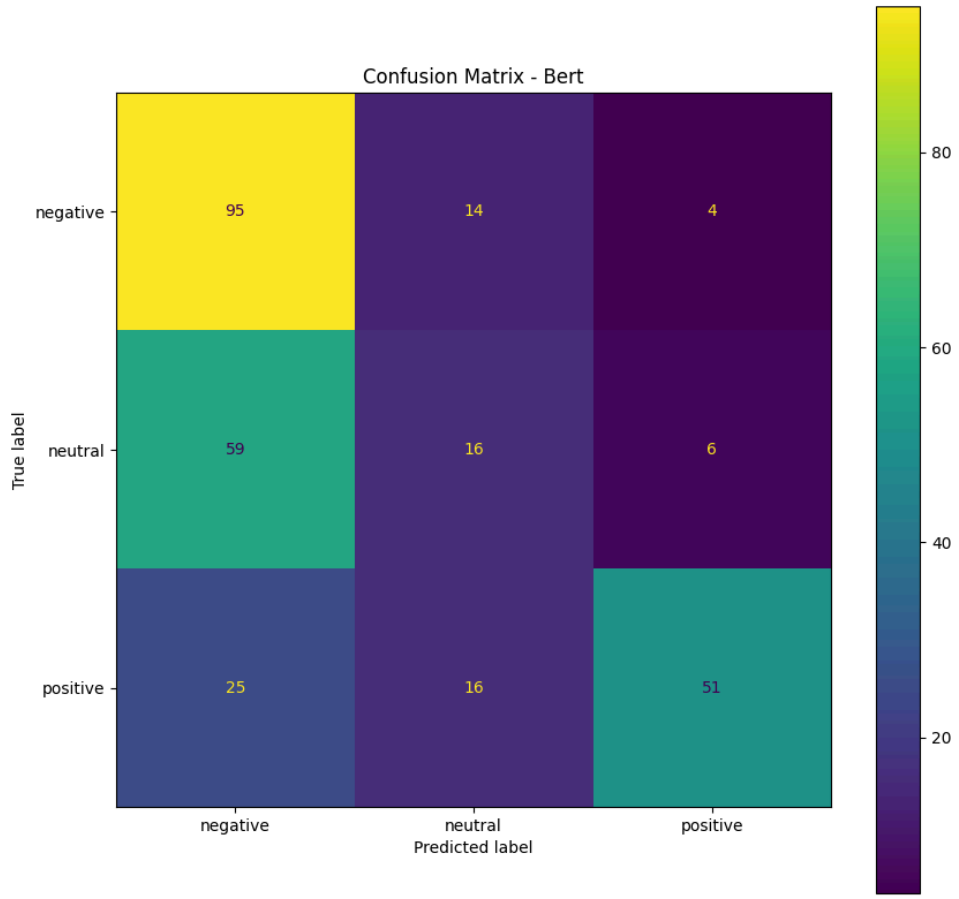
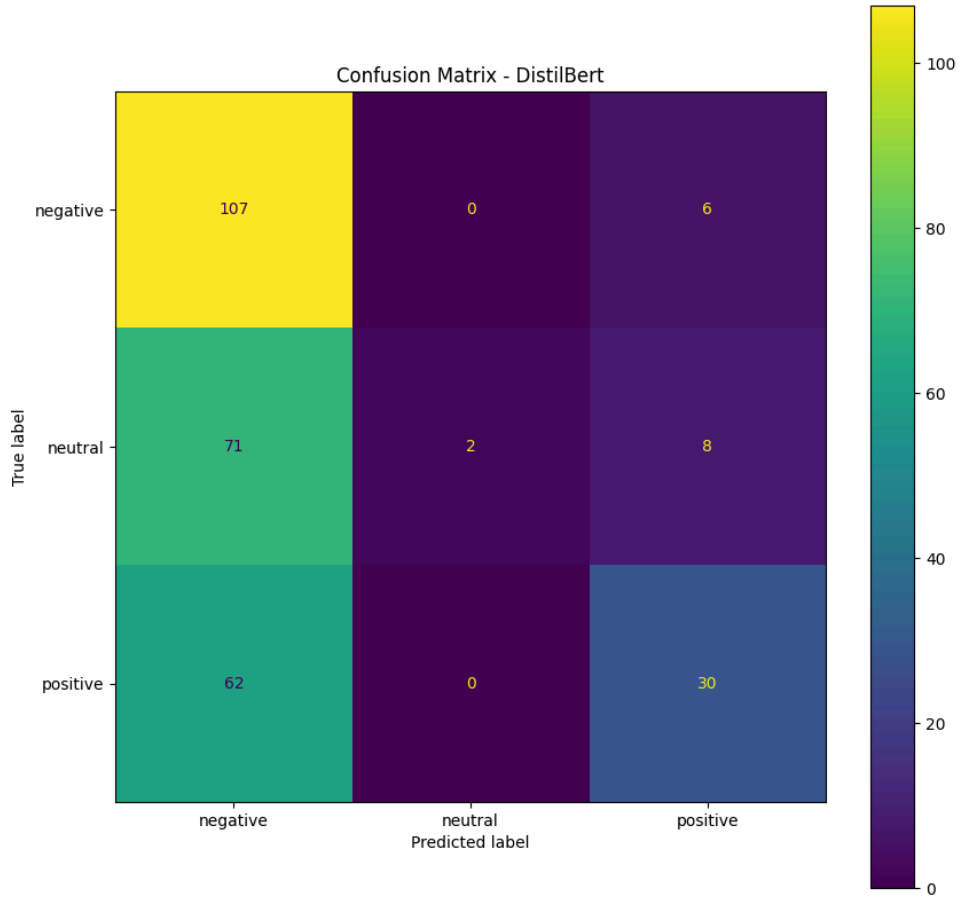


FIGURE 2
Confusion Matrix - DistilBERT



5. Final remarks, Limitations, and Future Work

This study analyzes the impact of the intensifier constructions *morrer de/a X* and *chorar de/a X* on sentiment analysis in Portuguese. Preliminary results show that metaphorical intensifier constructions significantly interfere with sentiment classification in the trained models available for Portuguese, mmBERT small multilingual sentiment and DistilBERT base multilingual cased sentiments. Both models struggled in the presence of MWEs with highly inconsistent labeling. The experiment presents limitations, as it evaluates only pre-trained, openly available models, without engaging in model training. Future work will involve training models with different architectures on the same dataset in order to provide a more robust and reliable comparison.

Acknowledgements

This study is funded by CAPES/DS (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/Coordination for the Improvement of Higher Education Personnel). Marcia dos Santos Machado Vieira thanks CNPq (National Council for Scientific and Technological Development, projects 409043/2021-4 and 312423/2022-5) and FAPERJ (Carlos Chagas Filho Foundation for Research Support in the State of Rio de Janeiro, project E-26/201.209/2022 (273339), SEI 260.003/003571/2022) for the research support.

REFERENCES

Augustyniak, Ł., Woźniak, S., Gruza, M., Gramacki, P., Rajda, K., Morzy, M., & Kajdanowicz, T. (2023). Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. arXiv. https://arxiv.org/abs/2306.07902

Brum, H., & Nunes, M. das G. V. (2018). Building a sentiment corpus of tweets in Brazilian Portuguese. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. xx–xx). European Language Resources Association. https://aclanthology.org/L18-1658

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). https://arxiv.org/abs/1810.04805

Johri, Prashant & Khatri, Sunil Kumar & Al-Taani, Ahmad & Sabharwal, Munich & Suvanov, Shakhzod & Chauhan, Avneesh. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. https:/10.1007/978-981-15-9712-1_31.

Kiritchenko, S., & Mohammad, S.M. (2016). The Effect of Negators, Modals, and Degree Adverbs on Sentiment Composition. *ArXiv, abs/1712.01794*.

Lakoff, G., & Johnson, M. Merophors we live by. Chicago: University of Chicago Press, 1980.

Mota, N.A., Nunes, L.F., dos Santos Machado-Vieira, M. (2021) Você vai ficar roxo de surpresa ao descobrir como intensificamos horrores. Roseta 4(1) , https://www.roseta.org.br/2021/01/29/voce-vai-ficar-roxo-de-surpresa-ao-descobrir-como-intensificamos-horrores/

Mozetič, I., Grčar, M., & Smailović, J. (2016). Multilingual Twitter sentiment classification: The role of human annotators. PLOS ONE, 11(5), e0155036. https://doi.org/10.1371/journal.pone.0155036

Taboada, M. (2016) Sentiment analysis: An overview from linguistics. Annual Review of Linguistics. 2: 325-347.

Silva, T.C.d.A. (2024): Explodir de tanto ódio: análise funcional-construcionista de [Xe-feito DE Yintensificador Zcausa]GRAD. Master's thesis, Universidade Federal do Rio Grande do Norte, Natal (2024), orientador: Dr. Edvaldo Balduino Bispo, Centro de Ciências Humanas, Letras e Artes.

Trott, Torrent, Chang & Schneider. (2020). (re)construing meaning in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5170–5184, Online. Association for Computational Linguistics.

Wankhade, M., Rao, A., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review 55, 5731–5780 (02 2022). https://doi.org/10.1007/s10462-022-10144-1