Building a corpus from supermarket reviews in portuguese for document-level Sentiment Analysis

Vinícius Takeo Friedrich Kuwaki
vtkwki@gmail.com
UDESC - Santa Catarina State University
Joinville, Santa Catarina, Brazil

Mateus Nepomuceno Ladeira
mateusnladeira@gmail.com
UDESC - Santa Catarina State University
Joinville, Santa Catarina, Brazil

Matias Giuliano Gutierrez Benitez
matiguti17@gmail.com
UDESC - Santa Catarina State University
Joinville, Santa Catarina, Brazil

Rui Jorge Tramontin Junior
tramontin@gmail.com
UDESC - Santa Catarina State University
Joinville, Santa Catarina, Brazil

ABSTRACT

Sentiment Analysis (SA) is a field of research within Natural Language Processing that has been growing in the last decades due to social media and smartphones popularization. Many SA applications make use of a corpus: a collection of data in textual form used to train and/or test SA resources. This work describes the construction of a corpus intended for document-level SA. The corpus contains reviews of supermarkets throughout Brazil, extracted from Google Places. The data were collected taking into account the Brazilian geographic distribution and linguistic variations, and were carefully reviewed. The corpus was then evaluated using a k-fold cross-validation method applied in both machine learning and deep learning techniques in which precision, accuracy, recall and f1-score were collected and compared among the techniques. It was also tested by a lexical approach using a domain specific lexicon.

KEYWORDS

Sentiment Analysis, Document Level, Corpus, Lexical Approach, Machine Learning, Deep Learning, Portuguese Language

1 INTRODUCTION

Social media and smartphones have brought a new context for data science, in special for Natural Language Processing (NLP), a subarea of Artificial Intelligence which uses spoken and written text as study subject. In this context, Sentiment Analysis (SA), a field within NLP, seeks to determine if a document containing an opinion (e.g., a product review) expresses a positive, neutral or negative polarity [26]. To accomplish this goal, several techniques can be applied, such as lexical, machine learning and deep learning approaches. Several of these techniques make use of a corpus: a collection of texts in electronic form that represents a language [41].

A corpus must represent a linguist variety as far as possible [41]. Considering this requirement, this paper presents a corpus built from reviews extracted from Google Places. The data were collected taking into account the vast territorial area of Brazil and its linguistic varieties. For the proposed corpus, data related to supermarket reviews were collected and manually reviewed. The corpus was then evaluated using the k-fold cross-validation method [7] for both machine learning and deep learning approaches.

In addition to the corpus construction, we also tested the use of a domain specific sentiment lexicon. A sentiment lexicon is a dictionary or a book of words containing their related polarity (positive, negative or neutral) [40]. A few sentiment lexicons for general purposes can be found in the literature, which means they are not for a specific domain. Wilson et. al [50] claims that words can have different meaning in different contexts. Regarding this, we built a sentiment lexicon specific for the domain of supermarket reviews. The aim is to test if there is any gain in performance in comparison with general purpose lexicons.

This paper is organized as follows. Section 2 presents the related work and an overview of SA concepts. Section 3 describes the methodology applied in the corpus construction and validation. The results are presented and discussed in Section 4. Final considerations and future work are presented in Section 5.

2 RELATED WORK AND OVERVIEW

SA is a field of research that have grown largely. Several studies have been conducted not only in computer science area, but also in medicine [15], psychology [49] and many others. Although English literature for SA, in special regarding corpora studies has been extensive, for Portuguese, there is still a large need [34].

2.1 Related Work

While English literature regarding SA corpora is extensive, Brazilian Portuguese still strives with a few corpora for SA studies. Ravi et. al [38] discuss several different *corpora* for English (and other languages), in which three of them [1, 23, 27] use/present a corpus with more than 11k texts. Also, great part of the *corpora* presented uses Twitter [1, 14, 22, 23] or reviews [3, 6, 7, 18, 27, 47, 48, 51] data from several different platforms and forums.

Portuguese Language has ReLI (REsenhas de LIvros) as its largest SA corpus: it contains 1600 reviews from 14 different books where the authors annotated 12.470 sentences taking their context into account [12]. Freitas and Vieira [9] have also presented an annotated corpus containing reviews in the accommodation sector from ten different TripAdvisor hotels. A corpus using tweets in Portuguese during the 2014 World Cup has also been built by [31], in which 2728 tweets were manually annotated and compiled into a single corpus. Considering the source for corpus construction, some works in English have already explored Google Places as a tool to collect location and/or reviews [25][32].

In the literature, there are a few sentiment lexicons for general purposes, which means they are not for a specific domain. LIWC-PT [2] and SentiLEX [5] are a few examples. On the other hand, several works have proposed the construction of domain specific lexicons in Portuguese, such as: stock market [35], telecommunications [11], small messages (such as tweets and SMS) [10][21].

2.2 Levels

As discussed earlier, SA goal is to determine if a text expresses a positive, neutral or negative opinion being that value a discrete or continuous value. For this work we used discrete values: -1, 0 and 1, corresponding to negative, neutral or positive, respectively.

SA can be performed in five different levels: document, sentence, aspect, concept and user level [42]. Document-level SA is the task of classifying a textual review, which is given on a single topic [30], accounting for its polarity, regarding its division in sentences. At sentence level, pieces of documents (that is, sentences) are analyzed individually. Aspect level, on the other hand, categorizes the sentiment according to specific aspects of entities present in the text [44]. Concept level is intended to infer the semantic and affective information associated with opinions to enable a comparative feature-based sentiment analysis [36]. Finally, user level SA analyzes what people think regarding a subject, rather than to quantify what is expressed in a sentence or a document [46]. This work focuses on the document level, hence the proposed corpus contains several reviews, which some of them are composed of one or many sentences and corresponds to several user's opinions about supermarkets throughout Brazil.

2.3 Techniques

SA techniques can be divided in machine learning, hybrid approaches, lexicon-based approaches, ontology-based and deep learning [42]. The corpus developed in this work was tested by techniques from the following approaches: lexicon-based, machine learning and deep learning.

Lexicon-based approaches focus on individual words, analyzing their frequency in a document. Such approaches calculate the document orientation based on the counting of the individual polarized words [16]. Many of these algorithms use a sentiment lexicon in order to gather the polarity of each word in the document, counting the number of negative, positive or neutral words and returning the maximum value among them. Some works have also considered the negative adverb "não" (not) as an inverter of a word's polarity [43, 45]. We tested an algorithm to count the number of positive and negative words present in a sentiment lexicon, inverting a words polarity every time that a negative adverb precedes it.

supervised learning algorithm (which means that it needs labeled data in order to learn how to classify)

Machine Learning (ML) is the ability of an algorithm to adapt to new circumstances and to detect and extrapolate patterns [39]. One of the tasks of ML is classification: algorithms that compute the probability of a given entity belonging or not to a given class. This kind of algorithm can be seen as a function $F(x) = y$ that computes the probability y of x belonging or not to a set of classes. For this work, we consider discrete values (-1,0 and 1) to representing the classes. The algorithms are trained to classify a document as belonging or not to one of these three classes, which means that in the end, a document is either negative (-1), neutral (0) or positive (1). We tested three different machine learning approaches: Logistic Regression, Naive Bayes and Support Vector Machine. All of the approaches used in this work are supervised algorithms, which means they use labeled data in order to make predictions.

Logistic Regression is a classifier which allows prediction of outcome variables by combination of continuous and discrete predictors [37]. The model takes a vector of characteristics and designates a weight for each column's value, estimating a multiple linear function. According to Equation 1 $f(D)$ is the estimate polarity of a document D, each b_i corresponds to the weights and each x_i corresponds to the presence of the predictor's value.

$$f(D) = b_0 + b_1 \times x_1 + ... + b_m \times x_m \quad (1)$$

Naive Bayes assumes that the probability of each word occurring in a document is independent of the occurrence of other words in a document [28]. By doing so, it calculates the probability based upon Bayes theorem (see Equation 2) in which a prior and a likelihood probability are calculated in order to determine if a document belongs or not to a class (positive, neutral or negative).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (2)$$

Support Vector Machine, in contrast to Logistic Regression and Naive Bayes, aims at finding the boundaries that separate clusters of data by taking a set of points and separating these points using mathematical formulas [20]. For a three class problem, such ours, the algorithm clusters the data into three different sections: positive, neutral and negative documents. It uses the previously mentioned mathematical formulas to classify new sentences by putting the document in the boundaries defined by the formulas and verifying the relation between the document to be classified and the clusters' boundaries.

Unlike the other machine learning techniques exposed before, Deep Learning applies artificial neural networks in order to learn tasks using multiple layers inspired by the biological brain. It is composed of several processing units, organized in layers, called neurons [24]. The algorithm can learn by adjusting the weights between neurons, reassembling the learning process, like biological brains do. Several types of structures and algorithms are explored in the literature. In this work, Multi-layer Perceptron (MLP), a algorithm that can learn a non-linear function approximation from a given set of features [8]. We tested three different type of activation functions: Sigmoid, Hyperbolic Tangent (Tanh) and Rectified Linear Unit (ReLU) [4].

3 METHODOLOGY

To build a corpus for Brazilian Portuguese, texts where collected using web scraping techniques[1] in Google Places, where "supermarkets" was the selected domain, since it is a type of establishment that even small cities have. After the data collection, a revision stage was executed and the final corpus was compiled at the end. In order to test the selected classification approaches, the corpus was

[1]Web scraping is the practice of gathering data through any means other than a program interacting with an API [29].

submitted to preprocessing and converted to a normal form and then applied to each tested classifier. A sentiment lexicon was also built specifically for the supermarket domain in order to increase the performance of the lexicon-based approaches.

3.1 Data collection

In order to collect data samples across Brazil, reviews were collect throughout the country, where the query string "Mercados em X" (can be translated to "Supermarkets in X") was applied to every X belonging to the three most populated cities of each one of the 26 states of Brazil [17]. Federal District, where the capital Brasilia is located, was also considered. Figure 1 presents the location of all the selected cities. The reviews are related to the top 20 establishments of each selected city. The texts are limited to 256 characters and their corresponding ratings (a value in 1-5 scale) were collected as well.

Figure 1: Selected cities distribution across Brazil's geography.

After the data collection, a total of 7483 sentences were obtained and submitted to human revision in order to remove noise sentences and to correct misspelling errors.

3.2 Data Revision

The review of the collected texts and their ratings was performed manually by four human annotators. In order to provide a standard for this process, the following guidelines were created:

(1) Correction of typographical errors (typos) and other misspellings;
(2) Correction of a rating that does not correspond to the sentiment expressed in the text;
(3) Normalization the ratings to $\{-1, 0, 1\}$ values;
(4) Removal of documents that do not express an opinion.

According to the first guideline, misspellings were corrected since they can increase the vocabulary[2] size and sometimes can even change the whole text context, e.g. in the sentence: "Mercado caro por causa da inflamação" ("Expensive market because of inflammation") it is clear that the author meant "inflation" and not "inflammation". Besides that, the most common issue consisted of typos, e.g. "comer" ("to eat") and "coemr" (typo)

In the second guideline, sentences with wrong ratings were corrected since they can dangerously affect the efficiency of the classifier, e.g. "muito bom" (very good) with 1/5 rating. In that case, since we cannot assume that it is an irony, a rating of 1 in a scale of -1,0,1 was given. In order to avoid human bias, we decided to only correct extreme cases (outliers) such as the ones presented here, in which the rating was 1/5 and the sentiment expressed in the text was positive, or when the rating was 5/5 and the sentiment expressed was negative.

The third guideline applies a normalizer in each review. The main purpose is to transform the original 1-5 scale into -1,0,1 values. To do that, we followed the Equation 3 only when not in conflict with the previous guideline.

$$\begin{cases} -1, & \text{if rating} \in \{1, 2\} \\ 0, & \text{if rating} = 3 \\ 1, & \text{if rating} \in \{4, 5\} \end{cases} \quad (3)$$

In texts in which multiple and opposite opinions are expressed, we prefer the dominant sentiment. Table 1 presents an example.

Table 1: Example of a text with multiple opposite sentiments.

PT-BR: Caixas mal educadas (-1). A comida estava fria (-1). Os preços são muito bons (1).
English: Rude attendants (-1). The food was cold (-1). The prices are very good (1).
Normalized polarity: -1, since there are two negative clauses and one positive.

Finally, according to the fourth guideline, sentences in which the text does not express an opinion were removed from the corpus, since they can difficult the classification process.

Further examples can be seen in Table 2. At the end, a total of 7121 texts where compiled in the final corpus, in which 1082 are negative, 1004 neutral and 5035 are positive.

3.3 Domain Specific Lexicon (DSL) Construction

In addition to the corpus used in the Machine and Deep Learning techniques, we developed a Domain Specific Lexicon (DSL). The idea is to compare its performance related to LIWC-PT [2], a general purpose lexicon. To do so, we extracted all verbs, adverbs and adjectives from the vocabulary and manually annotated them according to their polarity in the supermarket domain. For example, the word "espera" ("wait") can mean that the establishment has long lines, which indicates a negative feature. On the other hand, "espera" has a positive polarity provided by LIWC-PT, as in Portuguese this verb could also mean "to hope".

All these words were submitted to two human annotators which were responsible to determine their polarity (a number belonging to $\{-1, 0, 1\}$).

[2]Data structure containing all different words encountered in the corpus.

At the end, a total of 1995 words were annotated, in which 966 were verbs, 784 were adjectives, 245 were adverbs. Table 3 presents the distribution of the words considering the polarities for each category.

Table 2: Examples of treatments applied to some collected reviews.

Original text: Olá tudo bemmm!! **English:** Hi, how are you? **Original rating**: 5/5 **Normalized polarity**: - **Treatment**: (4) Removed document, as it does not express an opinion.
Original text: Produto **as** vezes sem preços. **Mais** tem bastante variedade nos produtos. **Corrected text**: Produto **às** vezes sem preços. **Mas** tem bastante variedade nos produtos. **English:** Products sometimes without pricing. But there is a large variety of products. **Original rating**: 3/5 **Normalized polarity**: 0 **Treatment**: (1) Correction of misspellings; (3) Normalization of the rating.
Original text: Ótimo atendimento vocês estão de parabéns (wink emoji) **English:** Great service, congratulations [wink emoji]. **Original rating**: 1/5 **Normalized polarity**: 1 **Treatment**: (2) The polarity was corrected to 1, as the review expressed a positive sentiment.

Table 3: Polarity distribution across the domain specific lexicon (DSL).

	Verbs		Adjectives		Adverbs	
POL	Total	%	Total	%	Total	%
-1	159	16.46	270	34.44	26	10.61
0	666	68.94	353	45.03	202	82.45
1	140	14.49	161	20.54	17	6.94
Total	966		784		245	

3.4 Preprocessing

The preprocessing step was conducted by three different approaches: one for lexical methods and two different for both machine and deep learning methods (called L and N). These approaches can be seen in Table 4.

For lexical approaches, only alphabetical tokens and punctuation marks were maintained. The whole text was converted to lowercase. For machine and deep learning methods, two approaches L and N were tested. The L approach is analogous to the one for lexical approaches, except that all words are lemmatized[3], all stop-words[4] were also removed. The N approach, on the other hand, maintains the stopwords and merges all negation tokens, e.g. "não gosto" ("do not like") becomes a special token "NOT_gosto".

3.5 Bag of Words (BoW)

After preprocessing, the data was normalized using the Bag of Words (BoW) approach. This approach consists of converting textual information into numerical input for the models to be trained and tested. A BoW is an unordered set of words in with their position is ignored, keeping only the frequency of its occurrences in the document [19]. The BoW approach consists first in creating a vocabulary with n words. Afterwards, for each document of the corpus, an array with size n is built, where each position corresponds to a different word in the vocabulary. Each element contains the value 1, if the word occurs in the document, or 0 otherwise.

3.6 Lexical Approach

In this approach, we count the number of positive and negative words for each sentence, inverting their polarity in case of finding a negation token (such as "not"). After calculating the polarities for all sentences, the summation of all values determines the resulting polarity. Algorithm 1 presents the main idea of the lexical approach.

3.7 Machine Learning and Deep Learning Approaches

In order to test the previously discussed Machine Learning and Deep Learning approaches, *scikit-learn package* (version 1.0.2) [33] for Python Language was used. The following subsections present the used parameters.

3.7.1 Logistic Regression. In this classifier, the following parameters were modified:

- **C**: 30;
- **class_weigth**: balanced;
- **solver**: newton-cg;
- **multi_class**: multinomial;
- The other parameters kept the default values of the package[5].

3.7.2 Naive Bayes. We used only the default parameters of the package[6]

3.7.3 Support Vector Machine. In this classifier, only the decision function shape was changed:

- **decision_function_shape**: ovo;
- The other parameters kept the default values of the package[7].

[3]Converting the inflected form of a word, returning its dictionary form also known as *lemma*. For instance the word *mice* should be converted to *mouse* after lemmatization.
[4]Words that do not add meaningful semantic information for text classification, such as *the*, *at*, *which*, *on*, among others.
[5]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[6]https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html.
[7]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Algorithm 1: Lexical Approach Pseudo-code

```
Function classify text,lexicon
    polarity ← 0;
    hasNot ← false;
    forall sentence in text do
        sentencePolarity ← 0;
        forall token in sentence do
            if token is a negation token then
                hasNot ← true;
            end
            else if token is in lexicon and it is positive then
                sentencePolarity ← sentencePolarity + 1;
            end
            else if token is in lexicon and it is negative then
                sentencePolarity ← sentencePolarity - 1;
            end
        end
        if hasNot is true then
            sentencePolarity ← sentencePolarity * -1;
            hasNot ← false;
        end
        polarity ← polarity + sentencePolarity;
    end
    if polarity > 0 then
        return 1;
    end
    else if polarity = 0 then
        return 0;
    end
    else
        return -1;
    end
end
```

3.7.4 Multi-layer Perceptron. For this classifier, we tested four different activation functions, as mentioned before, changing the following parameters:

- **solver**: adam;
- **alpha**: 1e-5;
- **hidden_layer_sizes**: (101,);
- **max_iter**: 900;
- The other parameters kept the default values of the package[8].

3.8 Validation Method

In order to test the approaches described previously, a k-fold cross-validation method [13] was used for both machine learning and deep learning approaches, considering a 10 fold size validation, in which accuracy, precision, f1-score and recall measurements were collected. The same measurements were also collected for lexical approaches, however without the cross-validation method, since a training stage is not needed.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Table 4: Preprocessing strategies used.

Approach	Applied Techniques
Lexical	- Converting all words to lowercase; - Maintaining only punctuation marks and alphabetical tokens.
L	- Converting all words to lowercase; - Maintaining only punctuation marks and alphabetical tokens; - Removing stopwords; - Lemmatization.
N	- Converting all words to lowercase; - Maintaining only punctuation marks and alphabetical tokens; - Merging "not" with the subsequent word, creating a new special token; - Lemmatization.

4 RESULTS

After the data revision and preprocessing, the data was submitted to the classifiers described before. Apart from the approach using the LIWC-PT lexicon, all other methods presented a good performance (around 80%), as discussed in the following subsections.

4.1 Lexicon-based Approaches

As expected, the algorithm based on LIWC-PT did not provide good results compared to the other tested approaches. This confirms the evidence already pointed out in the literature, that a domain specific dictionary achieves better results. In this case, the results were as good as the machine and deep learning approaches. Table 5 presents the results comparing the lexical methods.

4.2 Machine Learning Approaches

On opposite to lexical approaches, all machine learning methods have reached similar results. Naive Bayes, as expected, has performed better than the other algorithms in both L and N preprocessing strategies. Besides that, it is possible to see that the N preprocessing approach has improved classification on all three classifiers. Table 6 shows the collected measurements for machine learning algorithms.

4.3 Deep Learning Methods Results

Some of the methods tested have not performed better than the lexical approaches, but ReLU has achieved a better result than the algorithm based on DSL lexicon. Although the approaches have not performed better than Naive Bayes, they still have performed better than Logistic Regression.

Table 5: Lexical Methods Results.

	LIWC	DSL
Accuracy	0.560	**0.762**
Precision	0.667	**0.774**
Recall	0.560	**0.762**
F1	0.594	**0.767**

Table 6: Machine Learning Methods Results.

	Logist Regression		Naive Bayes		SVM	
	L	N	L	N	L	N
Accuracy	0.749	0.768	**0.815**	**0.825**	0.791	0.794
Precision	0.771	0.796	**0.793**	**0.804**	0.758	0.769
Recall	0.749	0.768	**0.815**	**0.825**	0.791	0.794
F1	0.759	0.779	**0.791**	**0.808**	0.759	0.768

Table 7: Deep Learning Methods Results.

	Sigmoid		Tanh		ReLU	
	L	N	L	N	L	N
Accuracy	0.746	0.764	0.743	0.764	0.766	**0.784**
Precision	0.753	0.768	0.750	0.765	0.765	**0.785**
Recall	0.746	0.764	0.743	0.764	0.766	**0.784**
F1	0.749	0.765	0.746	0.764	0.765	**0.784**

4.4 Discussion

All algorithms (except the one based on LIWC-PT lexicon) have performed well, with results ranging from 0.74 to 0.82 across all collected measurements. As our objective was to build a corpus, much of the efforts has been concentrated in the selection and the revision of the original texts. Thus, the results presented here are the first iteration of the experiments. Further tests are needed in order to improve the results. Other preprocessing strategies can be applied as well as different parameter configurations can be adjusted in the classification algorithms, specially in the deep learning techniques, which have a good potential for future experiments.

Even with few modifications on the default parameters of the machine learning algorithms tested in *scikit-learn*, the results obtained so far show that the corpus can be a good data source for other applications and studies. Further work can be done for optimizing those parameters, and thus improving the performance of the tested algorithms.

5 CONCLUSION

In the last years, there has been growing attention to Sentiment Analysis as a topic of research. Although corpora options for SA in English are wide, for Brazilian Portuguese, on the other hand, they are still scarce. This work has built a new corpus for SA in Portuguese language from Google Places reviews about supermarkets establishments in Brazil. The data was carefully revised and compiled into a corpus. Besides that, a Domain Specific Lexicon was also developed. All these lexical resources along with the source code of the conducted tests are publicly available[9]. Thus, this paper presented contributions that can be useful for both Machine Learning and Lexical approaches.

Three different SA approaches were tested, along with some different preprocessing methods. At the end, Naive Bayes had performed better, achieving next to 80% in all metrics. Something expected, as the algorithm has performed well in other English corpora. Lexical approaches have also achieved good results, proving that the construction of a domain specific sentiment lexicon, in this case, specific for this corpus, provides good results. Deep Learning algorithms, on the other hand, stay in the middle of lexical and machine learning approaches, but still delivered good results in comparison with already largely researched techniques such as Naive Bayes, Logistical Regression and Support Vector Machines.

Several issues were carefully dealt while reviewing the collected data. There was a concern with the maintenance of the linguistic variety, since texts from several regions of Brazil were collected. At the end we decided to not make large changes in the text, in order to preserve regional expressions, only misspellings and very common spelling errors in Portuguese were addressed, as discussed in the data review section.

Annotating is an exhausting and time-consuming work. We intend to study techniques to automatically identify and correct cases when a rating does not correspond to the sentiment expressed in the text. This would make possible to safely test unsupervised machine learning techniques on Google Places data, since this type of technique does not require human intervention.

In order to collect the data used in this corpus, an ongoing construction framework was used. This framework is already capable of collecting data from several cities across the world. We intend to explore new possibilities to collect data more sparsely around a country than we did in this work. As we divided the data collection around the most populated cities, we prioritize large urban centers. We think that a corpus can gain more linguistic variety from a better distribution across small cities far from the large urban centers, which means testing other criteria to select the cities, such as the geographical distance among them.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*. 30–38.
[2] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.
[3] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 440–447.
[4] N. Buduma and N. Locascio. 2017. *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*. O'Reilly Media. https://books.google.com.py/books?id=80glDwAAQBAJ
[5] Paula Carvalho and Mário J Silva. 2015. SentiLex-PT: Principais características e potencialidades. *Oslo Studies in Language* 7, 1 (2015).
[6] Helen Costa, Luiz HC Merschmann, Fabricio Barth, and Fabricio Benevenuto. 2014. Pollution, bad-mouthing, and local marketing: the underground of location-based social networks. *Information Sciences* 279 (2014), 123–137.
[7] Fermín L Cruz, José A Troyano, Fernando Enríquez, F Javier Ortega, and Carlos G Vallejo. 2013. 'Long autonomy or long delay?'The importance of domain in opinion mining. *Expert Systems with Applications* 40, 8 (2013), 3174–3184.

[9] https://github.com/takeofriedrich/google-places-for-sentiment-analysis

[8] Kia Dashtipour, Mandar Gogate, Ahsan Adeel, Cosimo Ieracitano, Hadi Larijani, and Amir Hussain. 2018. Exploiting Deep Learning for Persian Sentiment Analysis. In *Advances in Brain Inspired Cognitive Systems*, Jinchang Ren, Amir Hussain, Jiangbin Zheng, Cheng-Lin Liu, Bin Luo, Huimin Zhao, and Xinbo Zhao (Eds.). Springer International Publishing, Cham, 597–604.

[9] Larissa A De Freitas and Renata Vieira. 2015. Exploring resources for sentiment analysis in Portuguese language. In *2015 Brazilian conference on intelligent systems (BRACIS)*. IEEE, 152–156.

[10] Karine França de Souza, Moisés Henrique Ramos Pereira, and Daniel Hasan Dalip. [n. d.]. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. ([n. d.]).

[11] Ana Catarina Barbosa Forte. 2015. Análise de comentários de clientes com o auxílio a técnicas de Text Mining para determinar o nível de (in) satisfação. (2015).

[12] Cláudia Freitas, Eduardo Motta, Ruy L Milidiú, and Juliana César. 2014. Sparkling vampire... lol! annotating opinions in a book review corpus. *New language technologies and linguistic research: a two-way Road* (2014), 128–146.

[13] Jerome H Friedman. 2017. *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

[14] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[15] Yohan Bonescki Gumiel, Isabela Lee, Tayane Arantes Soares, Thiago Castro Ferreira, and Adriana Pagano. 2021. Sentiment Analysis in Portuguese Texts from Online Health Community Forums: Data, Model and Evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, 64–72.

[16] Fatemeh Hemmatian and Mohammad Karim Sohrabi. 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review* 52, 3 (2019), 1495–1545.

[17] IBGE. 2021. Estimativas da População residente no Brasil e unidades da federação.

[18] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 815–824.

[19] Daniel Jurafsky and James H Martin. 2017. Naive bayes and sentiment classification. *Speech and language processing* (2017), 74–91.

[20] Jayashri Khairnar and Mayura Kinikar. 2013. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications* 3, 6 (2013), 1–6.

[21] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50 (2014), 723–762.

[22] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. 2013. Ontology-based sentiment analysis of twitter posts. *Expert systems with applications* 40, 10 (2013), 4065–4074.

[23] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 538–541.

[24] Bing Liu Lei Zhang, Shuai Wang. 2018. Deep learning for sentiment analysis: A survey. (2018).

[25] Wei Lun Lim, Chiung Ching Ho, and Choo-Yee Ting. 2020. Tweet sentiment analysis using deep learning with nearby locations as features. In *Computational Science and Technology*. Springer, 291–299.

[26] Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 2, 2010 (2010), 627–666.

[27] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.

[28] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, Vol. 752. Citeseer, 41–48.

[29] Ryan Mitchell. 2018. *Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc.".

[30] Rodrigo Moraes, João Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications* 40, 2 (2013), 621–633.

[31] Silvia MW Moraes, Isabel H Manssour, and Milene S Silveira. 2015. 7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, 21–25.

[32] Javier Murga, Gianpierre Zapata, Heyul Chavez, Carlos Raymundo, Luis Rivera, Francisco Domínguez, Javier M Moguerza, and José María Álvarez. 2020. A sentiment analysis software framework for the support of business information architecture in the tourist sector. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLV*. Springer, 199–219.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[34] Denilson Alves Pereira. 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review* 54, 2 (2021), 1087–1115.

[35] Vitor Peres, Renata Vieira, and Rafael Bordini. 2019. Análises de Sentimentos: abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro. *Recuperado em de http://www. comp. ita. br/labsca/waiaf/papers/VitorPeres_paper_6. pdf* (2019).

[36] Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69 (2014), 45–63.

[37] Anjuman Prabhat and Vikas Khullar. 2017. Sentiment classification on big data using Naïve Bayes and logistic regression. In *2017 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–5.

[38] Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems* 89 (2015), 14–46.

[39] Stuart Russel and Petter Norvig. 2021. *Artificial Intelligence: A Modern Approach* (fourth ed.). Pearson Education.

[40] Dipanjan Sarkar. 2016. *Text Analytics with python*. Springer.

[41] John Sinclair. 2005. Corpus and text-basic principles. *Developing linguistic corpora: A guide to good practice* 92 (2005), 1–16.

[42] Rahul Kumar Singh, Manoj Kumar Sachan, and RB Patel. 2021. 360 degree view of cross-domain opinion classification: a survey. *Artificial Intelligence Review* 54, 2 (2021), 1385–1506.

[43] VK Singh, R Piryani, Ahsan Uddin, and P Waila. 2013. Sentiment analysis of Movie reviews and Blog posts. In *2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, 893–898.

[44] Hoong-Cheng Soong, Norazira Binti A Jalil, Ramesh Kumar Ayyasamy, and Rehan Akbar. 2019. The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques. In *2019 IEEE 9th symposium on computer applications & industrial electronics (ISCAIE)*. IEEE, 272–277.

[45] Marlo Souza and Renata Vieira. 2012. Sentiment analysis on twitter data for portuguese language. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 241–247.

[46] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1397–1405.

[47] Marilyn Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 812–817.

[48] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 783–792.

[49] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 201–213.

[50] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*. 347–354.

[51] Yunjie Calvin Xu, Cheng Zhang, and Ling Xue. 2013. WITHDRAWN: Measuring product susceptibility in online product review social network.