

# PELESent: Cross-domain polarity classification using distant supervision

Edilson A. Corrêa Jr, Vanessa Q. Marinho, Leandro B. dos Santos,  
Thales F. C. Bertaglia, Marcos V. Treviso, Henrico B. Brum  
Institute of Mathematics and Computer Science  
University of São Paulo (USP)  
São Carlos, São Paulo, Brazil

Email: {edilsonacjr, vanessaqm, leandrobs, thales.bertaglia, marcostreviso, henrico.brum}@usp.br

**Abstract**—The enormous amount of texts published daily by Internet users has fostered the development of methods to analyze this content in several natural language processing areas, such as sentiment analysis. The main goal of this task is to classify the polarity of a message. Even though many approaches have been proposed for sentiment analysis, some of the most successful ones rely on the availability of large annotated corpus, which is an expensive and time-consuming process. In recent years, distant supervision has been used to obtain larger datasets. So, inspired by these techniques, in this paper we extend such approaches to incorporate popular graphic symbols used in electronic messages, the *emojis*, in order to create a large sentiment corpus for Portuguese. Trained on almost one million tweets, several models were tested in both same domain and cross-domain corpora. Our methods obtained very competitive results in five annotated corpora from mixed domains (Twitter and product reviews), which proves the domain-independent property of such approach. In addition, our results suggest that the combination of emoticons and emojis is able to properly capture the sentiment of a message.

## I. INTRODUCTION

In the last few years, Sentiment Analysis has become a prominent field in natural language processing (NLP), mostly due to its direct application in several real-world scenarios [1], such as product reviews, government intelligence, and the prediction of the stock markets. One of the main tasks in Sentiment Analysis is the polarity classification, i.e., classifying texts into categories according to the emotions expressed on them. In general, the classes are positive, negative and neutral.

A popular application of polarity classification is in social media content. Microblogging and social networks websites, such as Twitter, have been used to express personal thoughts. According to Twitter’s website<sup>1</sup>, more than 500 million short messages, known as tweets, are posted each day. The analysis of this type of content is particularly challenging due to its specific language, which is mostly informal, with spelling errors, out of the vocabulary words, as well as the usage of emoticons and emojis to express ideas and sentiments.

Machine learning methods have been widely applied to polarity classification tasks in the context of social networks. This is particularly evident in shared tasks such as the SemEval

Sentiment Analysis tasks [2], [3], where these methods usually outperform lexical-based approaches. However, a major drawback of machine learning is its high dependency on large annotated corpora, and since manual annotation usually is time-consuming and expensive [4], many non-English languages lack this type of resource or, when existing, are very limited and specific, as it is the case for Portuguese.

In this paper, we adapt a distant supervision approach [5] to annotate a large number of tweets in Portuguese and use them to train state-of-the-art methods for polarity classification. We applied these methods in manually annotated corpora from the same domain (Twitter) and cross-domain (product reviews). The obtained results indicate that the proposed approach is well suited for both: same domain and cross-domain. Moreover, it is a powerful alternative to produce sentiment analysis corpora with less effort than manual annotation.

This paper is organized as follows. Section II gives a brief overview of some approaches for sentiment analysis and presents some works that have applied distant supervision to this task. Our approach is described in Section III. The evaluation corpora, machine learning algorithms and results are given in Section IV. Finally, our conclusions are drawn in Section V.

## II. RELATED WORK

Currently, methods devised to perform sentiment analysis and, more specifically, polarity classification range from machine learning to lexical-based approaches. While machine learning methods have proved useful in scenarios where a large amount of training data is available along with top quality NLP resources (such as taggers, parsers and others), they usually have low performance in opposite scenarios. Since most non-English languages face resource limitations, for example Portuguese, lexical-based approaches have become very popular. Some works following this line are [6]–[8].

Another alternative for languages with fewer resources is the use of hybrid systems, which combine machine learning and lexical-based methods. Avanço et al. [9] showed that this combination outperforms both individual approaches. This may imply that the development of better individual elements will lead to better results in the final combination.

<sup>1</sup><https://business.twitter.com/en/basics.html>

Machine learning approaches rely on document representations, normally vectorial ones with features like  $n$ -grams [1], a simple example is the bag-of-words model. Once a representation has been chosen, several classification methods are available, such as Support Vector Machines (SVM), Naive Bayes (NB), Maximum Entropy (MaxEnt), Conditional Random Fields (CRF), and ensembles of classifiers [3].

Apart from the traditional features, such as  $n$ -grams, some researchers have taken advantage of word embeddings, which are known to capture some linguistic properties, such as semantic and syntactic features. A well-known example of word embeddings is *Word2Vec* [10], [11]. Algebraic operations, such as sum or average, can be applied to convert word vectors into a sentence or document vector [12], [13]. However, this representation does not consider the order of the words in the sentence.

Paragraph vectors [14] (also known as *Doc2Vec*) can be understood as a generalization of *Word2Vec* for larger blocks of text, such as paragraphs or documents. This technique has obtained state-of-the-art results on sentiment analysis for two datasets of movie reviews [14]. The main goal of these dense representations is to predict the words in those blocks. Two models were proposed by Le and Mikolov [14], in which one of them accounts for the word order.

In addition, deep neural networks also consider the word order. Their methods have achieved good results in sentiment analysis, as shown in [15]–[17] and in the SemEval Sentiment Analysis Tasks [2], [3]. Nevertheless, these approaches need large datasets for training. Distant Supervision is a good alternative to obtain these datasets for the training/pre-training of deep neural networks [16], [18], [19].

Distant supervision is an alternative to create large datasets without the need of manual annotation. Some works have reported the use of emoticons as semantic indicator for sentiment [5], [18], [20], [21], while others use emoticons and hashtags for the same purpose [22], [23]. Go et al. [5], the first work to apply distant supervision to Twitter data, collected approximately 1.6 million of tweets containing positive and negative emoticons – e.g. “:)” and “:(” – equally distributed into two classes. They combined sets of features – unigrams, bigrams, part-of-speech (POS) tags – in order to train machine learning algorithms (NB, MaxEnt and SVM) and evaluate those in manually annotated datasets. The best accuracy was achieved using unigram and bigram as features for a MaxEnt classifier.

Severyn and Moschitti [18] used Distant Supervision to pre-train a Convolutional Neural Network (CNN). An architecture similar to the one proposed by Kim [17]. The network is composed of a first layer to convert words in dense vectors, following a single convolutional layer with a non-linear activation function, max pooling and soft-max. Deriu et al. [19] used a combination of 2 CNNs with a Random Forest classifier. However, this approach did not obtain improvements with distant supervision.

Despite the numerous studies and investigations of different techniques and methods for polarity classification, the problem

TABLE I  
ALL EMOTICONS USED TO REPRESENT EMOTION.

Positive	Negative
:) :-) :D =)	:( :-(

of relying on large annotated corpora remains open and the difficulty is intensified in non-English languages. In this paper, our contributions are the adapted framework for building polarity classification corpus to Portuguese, the built corpus itself and an evaluation on different state-of-the-art methods using this corpus, for same domain and cross-domain corpora.

### III. APPROACH

Following the approach of Go et al. [5], we initially collected a large amount of tweets in order to create the distant supervision corpus. Only tweets in Portuguese were crawled, and no specific queries were employed. In total, 41 million of tweets were collected.

After collecting the tweets, the next step was to split them into positive and negative classes. In order to do so, we used lists of emojis and emoticons selected according to the sentiment conveyed by them. Therefore, the polarity of a tweet is determined by the presence of emojis and emoticons in it – if it only contains positive ones (from the positive list), its polarity is assigned as positive. If a tweet contains both positive and negative elements, it is discarded since it is likely to be ambiguous. Following this idea, we used the same list of emoticons used by Go et al. [5], which is presented in Table I. Go et al. [5] did not use emojis, but these graphic symbols are also employed to convey ideas and sentiments [24]. In contrast to the small set of emoticons, there are hundreds of possible emojis. Therefore, we selected a representative list with positive and negative emojis. All the emojis conveying positive emotion are presented in Fig. 1. Fig. 2 illustrates the selected ones with negative emotion.

After filtering the tweets by the aforementioned criteria, we obtained a labeled corpus comprising 554,623 positive tweets and 425,444 negative ones. This corpus was used to train the machine learning methods. It is important to highlight that emojis and emoticons were removed from the tweets in the final corpus, so that their presence as a sentiment indicator is not learned by the models.

In addition to the filtering process, some preprocessing steps were performed to improve the corpus quality. Details about the preprocessing steps are given in the Supplementary Material, Section A. After these steps, tweets containing less than 4 tokens were discarded from the corpus. The complete framework (tweets collection, filtering and preprocessing methods) along with all experimental evaluation will be made available <sup>2</sup>.

<sup>2</sup><https://github.com/edilsonacjr/pelesent>

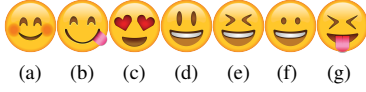


Fig. 1. All emojis used to represent positive emotion. Their respective unicones are: (a) U+1F60A, (b) U+1F60B, (c) U+1F60D, (d) U+1F603, (e) U+1F606, (f) U+1F600, and (g) U+1F61D.

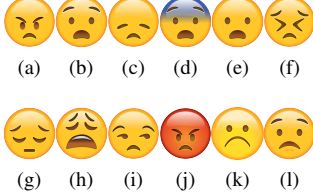


Fig. 2. All emojis used to represent negative emotion. Their respective unicones are: (a) U+1F620, (b) U+1F627, (c) U+1F61E, (d) U+1F628, (e) U+1F626, (f) U+1F623, (g) U+1F614, (h) U+1F629, (i) U+1F612, (j) U+1F621, (k) U+2639, and (l) U+1F61F.

#### IV. EXPERIMENTAL EVALUATION

In order to evaluate the quality of the corpus built using distant supervision, we trained state-of-the-art methods for polarity classification and applied the learned models to 5 well known manually annotated sentiment corpora. In the following, we present these corpora along with the message polarity classification methods, and finally, the obtained results.

##### A. Corpora

Sentiment classifiers are usually trained on manually annotated corpora. Because sentiments may be expressed differently in different domains [4], it is common to create domain-specific corpus. Since we intend to create a robust and generic corpus that is not domain-specific, we selected 5 corpora for evaluation, 2 being from the same domain (Twitter) and 3 from a different domain (product reviews). Below, we present the corpora that were used.

a) *Brazilian Presidential Election* [25]: This dataset is formed by tweets about the Brazilian presidential election run in 2010. The corpus is divided in two parts, one referencing Dilma Rousseff (*BPE-Dilma*) and the other José Serra (*BPE-Serra*), both being the most popular candidates in the election. The corpora were manually annotated in positive and negative, and used to evaluate stream based sentiment analysis systems.

b) *Buscapé* [26]: This dataset is formed by product reviews extracted from Buscapé website<sup>3</sup>. The documents were automatically labeled based on two informations given by the users. The first (*Buscape-1*) is based on a recommendation tag while the second (*Buscape-2*) is based on a 5-star scale (1-2 stars for negative and 4-5 stars for positive). Both corpora are balanced between the two classes, even though there is a notable difference on their sizes, possibly due to the low use of the recommendation tag.

<sup>3</sup><http://www.buscape.com.br/>

c) *Mercado Livre* [9]: Similar to the Buscapé dataset, this corpus is formed by product reviews from the online marketplace Mercado Livre<sup>4</sup>. The corpus was also automatically annotated based on a 5-star scale given by the authors of the reviews. The dataset is balanced between the positive and negative classes.

Table II presents a summary of the corpora.

TABLE II  
DATASETS USED IN THE EVALUATION OF THE SYSTEM.

Dataset	Total	Positive	Negative
BPE-Dilma	66,640	46,805	19,835
BPE-Serra	9,718	1,371	8,347
Buscape-1	2,000	1,000	1,000
Buscape-2	13,685	6,873	6,812
Mercado Livre	43,318	21,819	21,499

##### B. Machine Learning Methods

Machine learning has dominated the area of sentiment analysis, mostly because its high performance when manually annotated data is available. However, thanks to the great variety of methods, there is no consensus about which method is the best in this scenario. In the last editions of SemEval Sentiment Analysis Task, most of the best methods/systems used deep learning techniques [2], [3]. In this work, the evaluated methods range from simple linear models for classification using vector space models to hybrid (machine learning and lexical-based) and Deep Learning methods. The idea was to thoroughly evaluate the quality of the corpus regardless of the technique being used for learning. Below, each method is briefly described.

a) *Logistic Regression (LR)*: Also known as logit regression, LR can be understood as a generalization of linear regression models to the binary classification scenario, where a sigmoid function outputs the class probabilities [27]. In this paper, the logistic regression model predicts the class probabilities of a text, where the classes are "positive" and "negative". As input for this classifier, three text representations were used: (1) a bag-of-words model (LR+tfidf), where each document (tweet or review) is represented by its set of words weighted by *tf-idf* [28]; (2) a word embeddings based model (LR+w2v), where each document is represented by the weighted average of the embedding vectors (generated by Word2Vec [10], [11]) of the words that compose the document, the weights are defined by *tf-idf*; (3) the Paragraph Vector model (LR+d2v), which uses a neural network to generate embeddings for words and documents simultaneously in an unsupervised manner. Only the vectorial representations of documents were used by the classifier.

b) *Convolutional Neural Networks (CNNs)*: With the popularity of deep learning, CNNs have been applied to many different contexts, including several NLP tasks [29] and, more specifically, sentiment analysis [16]–[19]. Our CNN is similar

<sup>4</sup><http://www.mercadolivre.com.br/>

to the architecture proposed by Kim [17], which uses a single convolutional layer. In this architecture, the network receives as input a matrix representing the document, and each word in the document is represented by a dense continuous vector. The output of the network is the probability of a document being negative or positive.

*c) Recurrent Convolutional Neural Networks (RCNNs):*

This deep neural architecture uses both convolutional and recurrent layers. Recently explored by many works in NLP [30]–[32], this architecture has been successfully applied to sentiment analysis [3], [32]. Our architecture consists of a slight modification of the one used by Treviso et al. [31], where the final layer returns the probability for the whole document, indicating a positive/negative polarity. Using this combination of convolutional and recurrent layers, we explored the principle that nearby words have a greater influence in the classification, while distant words may also have some impact.

*d) Hybrid:* This method is a combination of two classifiers previously used for sentiment classification in cross-domain corpora [9] and follows the same setting introduced by Avanço [33]. The method consists of a SVM classifier combined with a lexical-based approach. The documents are represented by arrays of features including a binary bag-of-words (presence/absence of terms), emoticons, sentiment words and POS tags. Documents located near the separation hyperplane (in a threshold assumed as 0.5) learned by the SVM are considered to be uncertain. Those documents are then classified with a lexical-based approach, that uses linguistic rules for polarity classification in Portuguese.

For all methods, well-known machine learning libraries were used, such as Scikit-learn [34] and Keras [35]. Particularities such as parameters, details about the architecture, initializations and others can be found in the Supplementary Material, Section B.

### C. Results and Discussion

To evaluate and compare the methods in each corpus, F1 score (macro-averaged), recall (macro-averaged) and accuracy were chosen, mostly because of their traditional use in sentiment analysis [2], [3].

The main results are shown in Table III. Along with the results of each polarity classification method, we present the state-of-the-art (SotA) result reported for each corpus. Because the BPE corpora were conceived for a different context, there are no SotA reported results for those corpora. We also ranked each evaluated method by its F1 score.

The differences between the best method (in bold) and the SotA vary between 9.69% and 12.24%, very competitive results given the fact that all SotA reported results were obtained by a 10-fold cross validation scheme and our methods used a corpus from a different domain for training. Of all the methods, the Hybrid was the one that had the best performance in the corpora of product reviews. Such a result was due to the regularity of the language in this type of corpus, which makes lexical approaches highly effective. However, in domains such as Twitter, errors, abbreviations and slangs are very common,

which decreases the effectiveness of lexical-based approaches. This effect can be seen in the BPE-Dilma corpus.

An important aspect of Sentiment Analysis is the sensitivity of its methods to elements such as domain and temporality. In our evaluation, both were present in the selected corpora, which demonstrates the robustness of the constructed corpus and its resilience to temporality and the non-regularity of the language.

Regarding the deep learning methods (CNN and RCNN), both presented high rankings in almost all corpora. However, there was no huge difference between deep and shallow methods (logistic regression + document representation), indicating that large datasets decrease the performance difference between methods from different natures (even between simple and complex methods), a result commonly found in the big data era [36].

TABLE III  
RESULTS OBTAINED BY EACH METHOD TRAINED ON THE DISTANT SUPERVISION CORPUS.

Dataset	Method	F <sub>1</sub> score	Recall	Accuracy
BPE-Dilma	LR + w2v	0.5739 <sup>5</sup>	0.6037	0.5952
	LR + tfidf	<b>0.6477<sup>1</sup></b>	0.6443	<b>0.7128</b>
	LR + d2v	0.6135 <sup>4</sup>	0.6071	0.7256
	CNN	0.6337 <sup>3</sup>	0.6295	0.7051
	RCNN	0.6444 <sup>2</sup>	<b>0.6586</b>	0.6816
	Hybrid	0.5249 <sup>6</sup>	0.5855	0.5295
	SotA	—	—	—
BPE-Serra	LR + w2v	0.3515 <sup>6</sup>	0.4398	0.3915
	LR + tfidf	0.4110 <sup>5</sup>	0.5546	0.4475
	LR + d2v	0.5055 <sup>3</sup>	0.6028	0.5915
	CNN	0.4240 <sup>4</sup>	0.5929	0.4558
	RCNN	0.5286 <sup>2</sup>	0.5975	0.6426
	Hybrid	<b>0.5745<sup>1</sup></b>	<b>0.6073</b>	<b>0.7344</b>
	SotA	—	—	—
Buscape-1	LR + w2v	0.7232 <sup>4</sup>	0.7250	0.7250
	LR + tfidf	0.7469 <sup>3</sup>	0.7480	0.7480
	LR + d2v	0.6427 <sup>6</sup>	0.6465	0.6465
	CNN	0.6713 <sup>5</sup>	0.6870	0.6870
	RCNN	0.7654 <sup>2</sup>	0.7654	0.7654
	Hybrid	<b>0.7668<sup>1</sup></b>	<b>0.7695</b>	<b>0.7695</b>
	SotA	0.8892	—	0.8894
Buscape-2	LR + w2v	0.6814 <sup>6</sup>	0.6903	0.6910
	LR + tfidf	0.7725 <sup>2</sup>	0.7738	0.7742
	LR + d2v	0.7017 <sup>5</sup>	0.7027	0.7030
	CNN	0.7048 <sup>4</sup>	0.7115	0.7122
	RCNN	0.7656 <sup>3</sup>	0.7658	0.7657
	Hybrid	<b>0.7917<sup>1</sup></b>	<b>0.7930</b>	<b>0.7934</b>
	SotA	0.8935	—	0.8935
Mercado Livre	LR + w2v	0.6861 <sup>6</sup>	0.7048	0.7066
	LR + tfidf	0.8328 <sup>3</sup>	0.8329	0.8328
	LR + d2v	0.8089 <sup>4</sup>	0.8093	0.8097
	CNN	0.7745 <sup>5</sup>	0.7800	0.7813
	RCNN	0.8561 <sup>2</sup>	0.8561	0.8563
	Hybrid	<b>0.8614<sup>1</sup></b>	<b>0.8614</b>	<b>0.8614</b>
	SotA	0.9583	—	0.9583

## V. CONCLUSION AND FUTURE WORK

In recent years, the polarity classification task has drawn the attention of the scientific community, mainly due to its direct application in scenarios such as social media content and product reviews. Even though machine learning methods present themselves as high performance alternatives, they suffer from the need of a large amount of data during their training phases. In this paper, we adapted a distant supervision approach to build a large sentiment corpus for Portuguese. State-of-the-art methods were trained on this corpus and applied to 5 selected corpora, from same domain and different domain (cross-domain). Competitive results were obtained for all methods, although our best results did not outperform the best ones reported for the same corpora.

As future works, we intend to explore ways to improve the quality of the distant supervision corpus by applying techniques to remove outliers and tweets that do not convey any sentiment or convey the wrong sentiment. We also intend to modify this framework to make it able to represent the neutral class.

## ACKNOWLEDGMENT

E.A.C.J. acknowledges financial support from Google (Google Research Awards in Latin America grant) and CAPES (Coordination for the Improvement of Higher Education Personnel). V.Q.M. acknowledges financial support from FAPESP (grant no. 15/05676-8). L.B.S. acknowledges financial support from Google (Google Research Awards in Latin America grant) and CNPq (National Council for Scientific and Technological Development, Brazil). T.F.C.B., M.V.T., and H.B.B. acknowledge financial support from CNPq. In part of this work a GPU donated by NVIDIA Corporation was used.

## REFERENCES

- [1] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [2] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 451–463.
- [3] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016)*, pp. 1–18, 2016.
- [4] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 751–760.
- [5] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.
- [6] M. Souza, R. Vieira, D. Buseti, R. Chishman, I. M. Alves *et al.*, "Construction of a portuguese opinion lexicon from multiple resources," in *8th Brazilian Symposium in Information and Human Language Technology*, 2011, pp. 59–66.
- [7] P. P. Balage Filho, T. Pardo, and S. Aluísio, "An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis," in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, S. M. Aluísio and V. D. Feltrim, Eds. Fortaleza-CE, Brazil: Sociedade Brasileira de Computação, 21–23 Oct. 2013, pp. 215–219.
- [8] L. V. Avançaço and M. d. G. V. Nunes, "Lexicon-based sentiment analysis for reviews of products in brazilian portuguese," in *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*. IEEE, 2014, pp. 277–281.
- [9] L. V. Avançaço, H. B. Brum, and M. G. Nunes, "Improving opinion classifiers by combining different methods and resources," in *XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 2016, pp. 25–36.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [12] Y. Zhou, Z. Zhang, and M. Lan, "Ecnu at semeval-2016 task 4: An empirical investigation of traditional nlp features and word embedding features for sentence-level and topic-level sentiment analysis in twitter," *Proceedings of SemEval*, pp. 256–261, 2016.
- [13] E. A. Corrêa Jr, V. Q. Marinho, and L. B. d. Santos, "Nilc-usp at semeval-2017 task 4: A multi-view ensemble for twitter sentiment analysis," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*, 2017.
- [14] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, T. Jebara and E. P. Xing, Eds., 2014, pp. 1188–1196.
- [15] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642.
- [16] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 655–665. [Online]. Available: <http://www.aclweb.org/anthology/P14-1062>
- [17] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1746–1751.
- [18] A. Severyn and A. Moschitti, "Unitn: Training deep convolutional neural network for twitter sentiment classification," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 464–469.
- [19] J. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. De Luca, and M. Jaggi, "Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision," *Proceedings of SemEval*, pp. 1124–1128, 2016.
- [20] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL student research workshop*. Association for Computational Linguistics, 2005, pp. 43–48.
- [21] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREC*, vol. 10, no. 2010, 2010.
- [22] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd international conference on computational linguistics: posters*. Association for Computational Linguistics, 2010, pp. 241–249.
- [23] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg," in *In The International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social*. Association for the Advancement of Artificial Intelligence, 2011.
- [24] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PloS one*, vol. 10, no. 12, p. e0144296, 2015.
- [25] I. S. Silva, J. Gomide, A. Veloso, W. Meira Jr, and R. Ferreira, "Effective sentiment stream analysis with self-augmenting training and demand-driven projection," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 475–484.
- [26] N. S. Hartmann, L. V. Avançaço, P. P. Balage Filho, M. S. Duran, M. D. G. V. Nunes, T. A. S. Pardo, S. M. Aluísio *et al.*, "A large corpus of

product reviews in portuguese: Tackling out-of-vocabulary words,” in *International Conference on Language Resources and Evaluation*, 9th. European Language Resources Association-ELRA, 2014.

- [27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- [28] G. Salton, “Automatic text processing: The transformation, analysis, and retrieval of,” *Reading: Addison-Wesley*, 1989.
- [29] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.
- [30] N. Kalchbrenner and P. Blunsom, “Recurrent convolutional neural networks for discourse compositionality,” *arXiv preprint arXiv:1306.3584*, 2013.
- [31] M. V. Treviso, C. Shulby, and S. M. Aluísio, “Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 315–325.
- [32] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *AAAI*, vol. 333, 2015, pp. 2267–2273.
- [33] L. V. Avanço, “Sobre normalização e classificação de polaridade de textos opinativos na web,” Master’s thesis, Universidade de São Paulo, 2015.
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [35] F. Chollet *et al.*, “Keras,” <https://github.com/fchollet/keras>, 2015.
- [36] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.
- [37] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [38] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

## SUPPLEMENTARY MATERIAL

### A. Preprocessing

In order to properly tokenize and preprocess the tweets, the following steps were performed:

- Punctuation marks forming an emoticon were considered as a single token (e.g. :- ( and ;) )
- Sequences of consecutive emojis were split so that each emoji formed a single token
- Additional punctuation marks (not forming any emoticon) were removed
- Usernames (an @ symbol followed by up to 15 characters) were replaced by the tag *USERNAME*
- Hashtags (a # symbol followed by any sequence of characters) were replaced by the tag *HASHTAG*
- URLs were completely replaced by the tag *URL*
- Numbers, including dates, telephone numbers and currency values were replaced by the tag *NUMBER*
- Subsequent character repetitions were limited to 3 – i.e., all sequences of the same character were trimmed to fit the limit of 3

The following tweet is used to illustrate the preprocessing steps:

**Original:** *hj tenho aula de manhã, tarde e noite. das 8h ate 19h :(( #cansado*

**Preprocessed:** *hj tenho aula de manhã tarde e noite das NUMBER ate NUMBER :(( HASHTAG*

### B. Details about the machine learning methods

a) *Logistic Regression:* There is no additional information about this classifier.

b) *Convolutional Neural Networks:* The complete architecture is presented in Figure 3, where the input layer is a matrix composed by  $n$  input words, and each word is a  $d$  dimensionality real vector. The convolutional layer receives these vectors as input and it is responsible for the automatic extraction of  $n_f$  features depending on a sliding window of length  $h = \{3, 4, 5\}$ . The output from the convolutional layer is then passed to the max-over-time pooling layer, and the new extracted features are concatenated. This results in a large dimensional vector that is passed to a fully connected layer, where the softmax operation [27] is applied, returning the probability of a document being negative or positive. In the penultimate layer, we employed dropout with a constraint on l2-norms of the weight vectors to reduce the chance of overfitting [37].

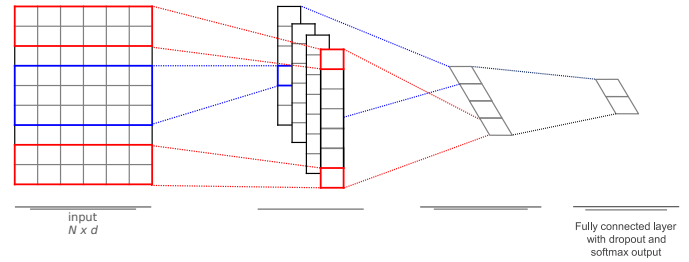


Fig. 3. CNN architecture adapted from [17].

c) *Recurrent Convolutional Neural Networks:* The complete architecture is illustrated in Figure 4. The architecture is composed by an input layer that has  $\varphi$  input features, and each feature has a dimensionality of  $d$ . The convolutional layer is responsible for the automatic extraction of  $n_f$  new features depending on 3 neighboring words. Then, a max-pooling operation is applied over time, looking at a region of  $h_m = 3$  elements to find the most significant features. The new extracted features are fed into a recurrent bidirectional layer which has  $n_f$  units known as Long Short-Term Memory [38], which are able to learn over long dependencies between words. Finally, the last recurrent state output is passed to a totally connected layer, where the softmax operation [27] is calculated, giving the probability of a document being negative or positive. Between these two layers, dropout is used to reduce the chance of overfitting [37].

For both neural network models (CNN and RCNN), we employed the early stopping strategy ( $p = 3$ ) to avoid overfitting, i.e. the training phase finishes when the validation loss has stopped improving. Other experimental settings for CNN and RCNN (number of epochs, batch size, learning rate, etc.) can be seen in their original papers [17], [31], respectively.

d) *Hybrid:* The two methods combined to classify a document in polarity classes are described below:

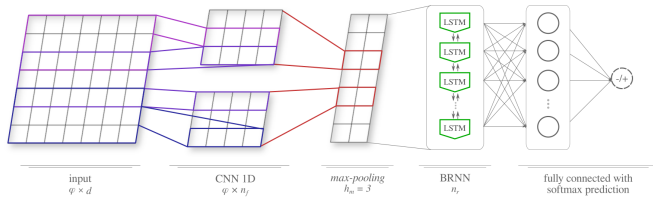


Fig. 4. RCNN architecture adapted from [31].

- SVM classifier: The SVM employed uses a RBF Kernel ( $\gamma = 1/n_{\text{features}}$ ),  $C$  defined as 0.25, and L1 penalty for regularization.
- Lexical-based classifier: Each word present in a sentiment lexicon receives a value according to its polarity. Positive words are valued as 1 and negative ones as  $-1$ . The presence of an intensification word (e.g. *muito*, *demais*) in a window around the word multiplies its value by 3. The presence of a downtoner divides the current value by 3. A negation multiplies the value of a word by  $-1$ , inverting its polarity. Whenever a negation is in the same window as an intensification, it becomes a downtoner (e.g. *não muito*), the same occurs with a downtoner (*não pouco*). The polarity values are then summed up to determine the document polarity.