Multi-label Classification of User Reactions in Online News

Zacarias Curi*, Alceu de Souza Britto Jr*† and Emerson Cabrera Paraiso*
*Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR, Brazil
{zacarias, alceu, paraiso}@ppgia.pucpr.br
† Universidade Estadual de Ponta Grossa (UEPG), Ponta Grossa, PR, Brazil

***Abstract*—The increase in the number of Internet users and the strong interaction brought by Web 2.0 made the Opinion Mining an important task in the area of natural language processing. Although several methods are capable of performing this task, few use multi-label classification, where there is a group of true labels for each example. This type of classification is useful for situations where the opinions are analyzed from the perspective of the reader. Recently, Deep Learning has been registering the state of the art in several single-label problems. This paper discuss the efficiency of the Long Short-Term Memory compared to traditional multi-label classification approaches. To do that, extensive tests were carried out on two news corpora written in Brazilian Portuguese annotated with reactions. A new corpus called BFRC-PT is presented. In the tests performed, the highest number of correct predictions was obtained with the Classifier Chains method combined with the Random Forest algorithm. When considering the class distribution, the best results were obtained with the Binary Relevance method combined with the LSTM and Random Forest algorithms.**

Keywords - multi-label classification; deep learning; LSTM; opinion mining.

I. Introduction

The success of Web 2.0 provides a constant generation of a large amount of textual data. The sites are very interactive. This characteristic combined with the cultural diversity of users ensures that different organizations are interested in the information contained in those texts. In this scenario, the task of opinion mining became popular in the area of Natural Language Processing (NLP) since it can provide the tools for information extraction and knowledge acquisition. Among the existing techniques, Deep Learning has been achieving good results for the classification task in cases where there is a large amount of data. An example of this is the application of the algorithm Long Short-Term Memory (LSTM) in the analysis of texts generated by the Web 2.0 published in [1] and [2].

Most of the opinion mining research perform the single-label classification, in which only one label is considered for each text. This type of classification is efficient in cases where the purpose is to analyze the opinion expressed by the writer. However, there are situations where the goal is to analyze the text from the reader's perspective, that is how the reader reacted when was reading the text. The term reaction is defined in this work as the attitude or sensation acquired by a person upon receiving a stimulus from an external source. The reaction can be presented as a component of the emotions. Desmet [3] defines that emotions can be treated as a multifaceted phenomenon, consisting of behavioural reactions, expressive reactions, physiological reactions and subjective feelings. In this work, we analyze a corpus annotated with expressive reactions and with emotions. As the corpora used are annotated from the perspective of the reader, it is necessary to use the multi-label classification, in which several reactions are considered simultaneously for the same text. This type of classification is necessary because each person has their individuality, which generates different reactions and consequently different emotions.

The task of multi-label classification can be accomplished through an adaptation of the classification algorithm or a transformation in the problem. The algorithm adaptation methods consider performing transformations in the traditional single-label classification methods to allow the use of multi-label problems. Problem transformation methods consider to transform a given problem into one or more single label problems [4].

This work aims to perform the task of multi-label classification of reactions in texts through the creation of several binary LSTM classifiers. This way, we intent to verify if the use of the LSTM allows better results than the traditional classification methods, as well as those obtained in several works with single-label classification.

This paper presents three main contributions. To the best of our knowledge, this is the first work to apply the LSTM algorithm with a problem transformation method for the task of classifying reactions in texts. Another contribution of this work is the introduction of a new corpus of online news written in Portuguese, labeled with user reactions. Finally, this work also evaluates some traditional methods of problem transformation considering different induction algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF).

The remainder of this paper is organized as follows. Section II presents some related works. Section III describes the methods and algorithms evaluated. The experiments and corresponding results are presented in Section IV and V, respectively. Finally, in Section VI we present our conclusions and future work.

II. Related Works

Most of the opinion mining works perform the single label classification. However, Liu and Chen [5] present the analysis of texts extracted from a Chinese microblog annotated in multi-label form. The authors presented a comparison with 11 methods of problem transformation and algorithm adaptation to classify these texts. Another way to accomplish this task is presented by Song and colleagues [6], in which lexicons were used.

Another work using lexicons is presented by Phan, Shindo and Matsumoto [7]. The authors report the creation of a new resource using a Recurrent Neural Network. The feature created is used for multi-label classification of Plutchik's basic emotions in transcripts of film dialogues. An approach using Deep Learning with a problem transformation technique is presented by Wang, Ren and Miao [8]. They proposed a method based on Convolutional Neural Network (CNN) for the multi-label classification of emotions in sentences of microblogs in Chinese.

Most opinion mining works use copora annotated from the writer's perspective. In the paper presented by Pool and Nissim [9], the authors use a corpus annotated from the perspective of the reader, using Facebook[1] messaging reactions. Although this work performs a single label classification, the used corpus contains more than one emotion associated with each text. Bhowmick and colleagues [10] used an algorithm adaptation technique called MLkNN to classify four emotions into a news corpus labeled from the perspective of the reader. Zhang et al. [11] present a new framework for classifying a corpus from the same perspective.

As for the single-label classification, most of the works existing in the literature perform the classification of texts in English or Chinese. Zwaan and colleagues [12] present the use of the Problem Transformation methods BR and RAkEL with the SVM algorithm for the classification of texts in Dutch. Another explored language is Japanese. Duan and colleagues [13] report the use of crowdsourcing for annotating two children's stories in that language. The authors also present two techniques based on the Problem Transformation methods called BR and LP with the NB classification algorithm. In this work, we use a corpus in Brazilian Portuguese annotated from the perspective of the reader. The novelty is related to the use of Deep Learning with a Problem Transformation Method for classification. The LSTM and the Problem Transformation methods used in this work are presented in the next section.

III. Classification Methods

This section briefly presents the multi-label classification methods used in the experimental part of this work. It also present the LSTM algorithm.

A. Multi-label classification

The most common approaches to traditional supervised learning tasks perform single-label classification. In this type of classification, each sample is represented by only one label. Considering λ as a single label for an instance of the database used and L as the class set of the problem, we have the classification called binary for cases where $|L| = 2$. In cases where $|L| > 2$ the classification is called multi-class. Different from binary and multi-class classification, where there is only one label λ for each instance, the multi-label classification accepts a set of labels Y to represent each instance, such that $Y \subseteq L$ [14]. In short, multi-label problems can be defined as situations where there is a set of true labels for each instance of the problem, and for at least one instance the set has more than one label. Currently, two groups of methods to solve this type of problem can be found in the literature: Problem Transformation and Algorithm Adaptation [4].

The Problem Transformation techniques consist of transforming the multi-label classification problem into one or more single-label sorting or classification problems. One way to accomplish this task is to create an independent binary classifier for each label of the problem, using the method called Binary Relevance (BR). The main problem of the BR method is that it does not consider the dependency between the labels, thus ignoring some characteristics of the problem [15]. One way to solve this is through the Classifier Chains (CC) method. The CC uses the output of a binary classifier as an input attribute to the next, thereby creating a link between binary classifiers and adding the relationship between classes in problem resolution [16]. The main problem of this method is the choice of the best order of the classifiers.

An alternative to the transformation of the multi-label problem into several binary problems, as performed in the BR and CC methods, is the transformation into a multiclass problem. An example of this is the Label Powerset (LP) method. This method creates a new label for each label combination in the training database. The main advantage of the strategy used by the LP method is the need for only one classifier. By contrast, the LP method can generate many new classes depending on the characteristics of the database used [17].

One way to reduce the problem of generating new classes in the LP method is to create class groups through a class ensemble method. One of these methods is the Random k-Labelsets (RAkEL), which creates a class ensemble for the LP method. The RAkEL method divides the initial set of labels into m random subsets with k classes called label sets. After this division, the label Powerset method is used to perform the transformation of the problem and enable the training [17]. An ensemble is also performed by the Hierarchy Of Multilabel classifiers (HOMER), where the multi-label problem is transformed into a hierarchical problem. The main advantage of the hierarchical division created is the use of fewer classes in each classifier and the more balanced distribution between these classes [18].

Besides the division of the multi-label problem into one or more classification problems, it is possible to carry out the transformation of the multi-label problem into a ranking problem. One of the methods that use this strategy is the Calibrated Label Ranking (CLR), introduced by [19]. The basic

[1] https://www.facebook.com/

idea of this method is to transform the multi-label problem into a label ranking problem, where the position of the labels is decided on the basis of peer-comparison techniques and is used to perform the classification. For this, the traditional Label Ranking algorithm is used to perform the ordering of the labels based on their relevance, after the ordering a calibration of labels is added, allowing the separation of the relevant labels from the irrelevant labels.

An alternative to problem transformation methods are the algorithm adaptation methods. These methods are defined as all traditional data mining algorithms that are adapted to work directly with a multi-label problem [4]. One of these changes is MLkNN, presented by [20] and [21]. In these works, the authors perform an adaptation of the KNN algorithm to allow the use of multi-label data. In the first step of the MLkNN algorithm, all the k nearest neighbors of each instance are identified. After this identification, statistical information obtained from the neighbor's label sets is used for use with the maximum a posteriori principle, which is used to determine the labels. The MLkNN algorithm and the other methods of algorithm adaptation in the literature are directly linked to its origin algorithm. Unlike these methods, problem transformation methods can be used with any single-label algorithm. The following section presents one of these algorithms, the LSTM.

B. Long Short-Term Memory

The Long Short-Term Memory (LSTM) algorithm, initially presented in [22], is a type of Recurrent Neural Network (RNN) capable of using long-term stored information for training. In conventional RNNs, it is possible to make a connection with some previous information, but this algorithm is unable to deal with distant information. To solve the problem of the dependence of terms, the model establishes a new structure called memory cell, shown in Figure 1. This structure is composed of an input gate, neurons with recurrent connections, a forget gate and an output gate. The first step of the LSTM is to use a sigmoid layer to decide what information will be discarded from the current cell. After choosing what will be discarded another sigmoid layer is used to decide what new information should be stored in the cell. Then a $tanh$ selects candidates to be stored in a vector. After the vector creation, a sigmoid function is used to decide which information is best for the next cell. With all the steps operated the old cell is updated and the process is performed again with the new data.

The steps performed by the LSTM can also be represented by the Equations 1 to 5. In these equations, the subscript characters represent vectors and the characters in uppercase arrays. In the notation used, f represents the forget gate, i represents the input gate, o the output gate, c represents the memory cell and h the LSTM unit. The matrices are W, which stores the input weights and U, which stores the recurring connections.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$

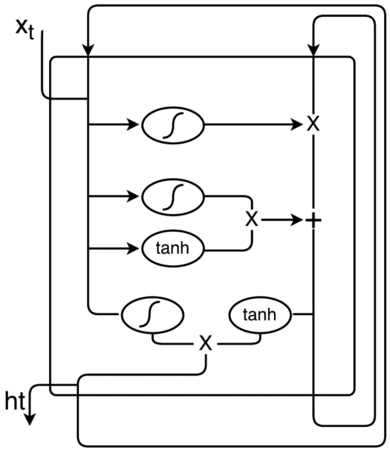

Fig. 1. Structure of a Long Short-Term Memory cell.

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{4}$$

$$h_t = o_t \circ \sigma_h(c_t) \tag{5}$$

The next section gives the details concerning the experiments performed in this work.

IV. Experiments

This section presents the experimental protocol followed during experimentation.

A. Portuguese news corpora

To evaluate our approach, two news corpora were used. The corpus called G1 was initially presented by Dosciatti and colleagues in [23]. This corpus is composed of 2,000 titles and headlines of news extracted from the website G1[2]. The news is annotated with the six basic emotions presented by Ekman [24] and the neutral class for cases where none of the emotions were present in the document. The classes used were: anger, disgust, fear, happiness, sadness, surprise and neutral. Originally, each news was labeled by two annotators, where each annotator identified the primary and secondary emotion of each news. In cases of a tie, a third annotator was consulted to define the primary emotion. The version of the corpus used in this work considers all the emotions selected by the annotators, being able to simultaneously have up to four labels. The number of examples of each label is shown in Figure 2. As can be seen in this figure, the G1 corpus is unbalanced, with 192 examples for the minority class (anger) and 848 for the majority class (sadness). The label cardinality of this corpus is 1.964 and the label density is 0.280.

In addition to the G1 corpus, we are presenting a new corpus named BuzzFeed Reactions Corpus (BFRC-PT) consisting in

[2] http://g1.globo.com/

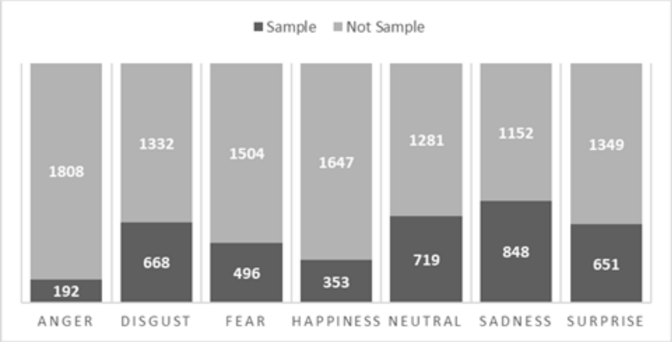

Fig. 2. Class Distribution of the G1 corpus

8,080 entertainment news written in Brazilian Portuguese, collected from the Brazilian version of BuzzFeed[3]. The news was annotated with the vote of the users. During the corpus collection (the first quarter of 2017), the site provided a field for the users to express their reactions for each news read. The eight labels of the presented corpus are defined based on these reactions, being: cute, fail, funny, hate, love, shock, skeptic and win. Because BuzzFeed is focused on entertainment, many news articles feature only pictures or videos, with no textual information. As the focus is the text, these news were discarded. Another change was the application of a threshold to discard the labels with few votes. Analyzing the votes, it was possible to observe inconsistencies, especially in the most popular news. For this reason, all labels with less than 3% of the total sum of the news labels were deleted. Even with the application of this threshold, the BFRC-PT has some degree of imbalance. Figure 3 presents the distribution of the labels of this corpus. The label cardinality of this corpus is 3.861 with the label density of 0.483. The corpus BFRC-PT can be accessed at link [4].

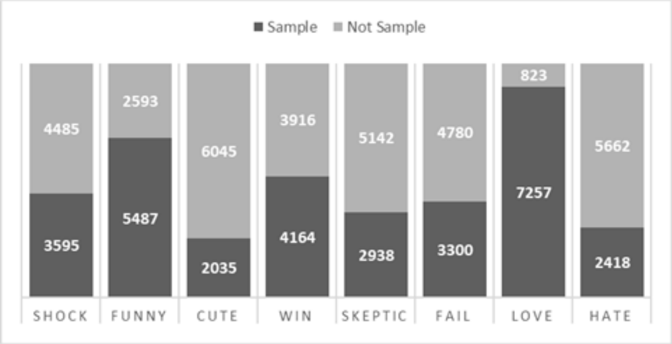

Fig. 3. Class distribution of the BFRC-PT corpus

B. Methods and Algorithms

The aim of this work is to verify the efficiency of the LSTM algorithm when used with the Binary Relevance method in comparison to the traditional approaches in the task of multi-label classification of reactions in texts. The Problem Transformation methods BR, CC, CLR, HOMER, LP and RAkEL were used with the NB, RF and SVM algorithms to be compared with the BR method with the LSTM algorithm and the adapted algorithm MLkNN. The parameters of the HOMER were defined based on the work [18]. The RAkEL has been tested with all available settings. All other methods have been tested with their default configuration. The experiments were performed with the implementations available in the Meka[5] and Mulan[6] software and with an implementation of the LSTM algorithm in the TensorFlow[7] framework. Due to unbalance, the databases were divided with 3-folds cross-validation. This division allows a greater number of examples of the minority classes in the test database, allowing an improvement in the evaluation.

For the pre-processing of the data used with the traditional algorithms, all texts were converted to the lowercase and special characters were removed. All words found in the stopwords list provided by SnollBall[8] were removed. A stemming, also provided by the snowball system, was applied to extract the radicals from the words. All links, emails, numbers, currency symbols and percentages were replaced by tokens. Finally, the TF-IDF (term frequency-inverse document frequency) method was applied to represent the words in vector form. For the LSTM algorithm, no changes were made to the words. For this algorithm, we used the embedding method word2vec with the vector pre-trained by Hartmann and Colleagues [25].

C. Evaluation Metrics

For the analysis of the used methods, two multi-label evaluation metrics were used: the hamming loss and the micro-F1. The hamming loss metric is defined by the Equation 6, where $\triangle$ implies the symmetric difference between two sets, X represents the test set, L the problem classes, $h(x_i)$ is the classifier prediction for the instance x_i, while Y_i corresponds to its label.

$$HL(h) = \frac{1}{|X|}\frac{1}{|L|}\sum_{i=1}^{|X|}|h(x_i) \triangle Y_i| \tag{6}$$

The metric F1$_{ml}$ constitutes the adaptation of the existing metric for single-label problems to multi-label problems. Like the original metric, the F1$_{ml}$ represents the harmonic mean between precision and recall being efficient to measure cases where the database is unbalanced. The adaptation occurs in the way the predicted values are calculated, when the values of each label are summed in Equation 7. After the definition of the values of the confusion matrix, the traditional F^{beta} metric is applied. This metric is presented in the Equation 8.

[3] https://www.buzzfeed.com/?country=pt-br

[4] https://www.ppgia.pucpr.br/∼paraiso/mineracaodeemocoes/recursos.php

[5] http://waikato.github.io/meka/

[6] http://mulan.sourceforge.net/

[7] https://www.tensorflow.org/

[8] http://snowballstem.org/

TABLE I
RESULTS OBTAINED

Classifier	Method	micro F1		hamming loss	
		G1	BFRC-PT	G1	BFRC-PT
KNN	**MLkNN**	0.46045	0.63438	0.31458	0.35809
NB	**BR**	0.52529	0.60962	0.26621	0.37651
	CC	0.54493	0.61479	0.27793	0.37953
	CLR	0.54070	0.63427	0.28243	0.37361
	HOMER	0.48756	0.60762	0.28749	0.37851
	LP	0.44118	0.61481	0.31343	0.37143
	RAkEL	0.54506	0.63103	0.27700	0.36465
RF	**BR**	0.56057	**0.65713**	0.25092	0.33685
	CC	0.50541	0.61007	**0.22678**	**0.33048**
	CLR	0.55483	0.65506	0.25249	0.34164
	HOMER	0.33531	0.59887	0.38022	0.39384
	LP	0.26623	0.37311	0.43635	0.61094
	RAkEL	0.55283	0.65243	0.25228	0.34672
SVM	**BR**	0.49096	0.61679	0.28507	0.36258
	CC	0.49120	0.61793	0.28621	0.36318
	CLR	0.53399	0.64679	0.27885	0.36154
	HOMER	0.22394	0.62281	0.54621	0.46839
	LP	0.28172	0.36706	0.74164	0.64954
	RAkEL	0.54850	0.64805	0.30835	0.35469
LSTM	**BR**	**0.56071**	0.6463	0.25416	0.35981

$$B_{micro}(h) = B\left(\sum_{j=1}^{|L|} VP_j, \sum_{j=1}^{|L|} FP_j, \sum_{j=1}^{|L|} VN_j, \sum_{j=1}^{|L|} FN_j\right) \tag{7}$$

$$F^{\beta}(h) = \frac{(1+\beta^2) \,.\, VP_j}{(1+\beta^2) \,.\, VP_j + \beta^2 \,.\, FN_j + FP_j} \tag{8}$$

V. Results and Discussion

The corpora used have different amounts of examples and texts with different sizes. These differences in the characteristics of the corpora can generate differences in the results of the classification methods. Table I presents the results obtained for both corpora.

As can be seen in Table I, the best result with corpus G1 for the micro F1 metric was established by the BR method with the LSTM algorithm. For the Hamming Loss metric, the best result was obtained by the CC problem transformation method with the RF classification algorithm. Random Forest also enabled the third, fourth and fifth best micro F1 when combined with problem transformation methods BR, CLR and RAkEL, respectively.

Although the RAkEL method obtained the third highest micro F1 for the tests performed with the RF algorithm in corpus G1, this method allowed the best results when combined with the SVM and NB algorithms. This method was tested with four different configurations for each classification algorithm used. For the RF algorithm, the best result was obtained with the creation of 14 subsets with 3 classes. For the NB algorithm, 10 subsets of 4 classes were created. The best result for the SVM algorithm was also obtained with the use of 10 subsets, but with 3 classes. The need for parameter settings to obtain the best result also occurred for the LSTM algorithm.

For the corpus G1, the best configuration of the LSTM algorithm was obtained using the first 50 words of the news represented in a 300-dimension embedding vector. The best configuration of the network has 25 neurons with a batch size of 200. The training was performed with 25 epochs with the Adam Optimizer and a learning rate of 0.01. In addition to the G1 corpus, the new corpus BFRC-PT was used. The best result was obtained with the use of the first 25 words of each news represented in an embedding matrix of 300 elements. The best configuration of the network for this corpus has 40 neurons with a batch size of 150. The training was carried out with 6 epochs and with the same function of optimization used for the corpus G1.

The differences between the corpora used generated differences in the parameters used and in the results obtained. For the BFRC-PT corpus, the BR method with the LSTM algorithm obtained a lower micro F1 to the same method with the RF algorithm. The LSTM algorithm was also inferior to the CLR and RAkEL methods with the RF and SVM algorithms. Although the values obtained were lower, the t test with confidence of 95% showed that there is no statistical difference between the results obtained with the LSTM and RF algorithms. In relation to the metric hamming loss, it is possible to observe that, as for corpus G1, the best result was obtained with the CC method in conjunction with the RF algorithm. The t test had showed that for this metric there is no statistical difference between the BR method with LSTM and the best result. Although the BR method with the LSTM algorithm obtained a lower result than the RAkEL and CLR methods with the SVM algorithm for the micro F1 metric, a higher result was recorded for the hamming loss metric. This result represents that although the LSTM obtained more correct predictions than the SVM, the distribution of the correct predictions among the classes was smaller.

The third and fourth best results for the micro F1 in BFRC-PT were obtained by the RAkEL method. As for the G1 corpus, different configurations of this method were tested for each algorithm used. For the RF and SVM algorithms, the best results were obtained with the use of 14 subsets with 3 classes. The best result for the NB algorithm was obtained with the use of 10 subsets of 4 classes. Unlike the RAkEL method, where different configurations were evaluated, the other methods tested were used with their default configurations or indicated settings. Among these methods are HOMER and LP, which generated the lowest results for the two corpora tested. These results demonstrate that the characteristics of corpora used make these methods less efficient.

VI. Conclusion

In this work, we presented the use of a Deep Learning algorithm with a problem transformation method for the multi-label opinion mining task. The LSTM classification algorithm was used by transforming the multi-label database into several binary databases using the BR method. For the evaluation of this technique was introduced a new corpus of news, labeled with user reactions. The two corpora used are composed of news in Brazilian Portuguese. For the comparison of the results obtained with the proposed method and with the methods established in the literature, tests with BR, CC, CLR, HOMER, LP and RA*k*EL were performed with NB, RF and SVM algorithms and with the algorithm adaptation method ML*k*NN.

The tests performed with G1 corpus demonstrated that the combination of the LSTM algorithm with the BR method allowed the highest micro F1 among all the evaluated methods. Although this combination was most efficient, there was a difference of only 0.014pp. between the result obtained by RF using the same method. For the hamming loss metric, the best result for the two corpora was obtained with the CC method with the RF algorithm. Although the best result for the metric hamming loss was the same for both corpora, the best result obtained by the micro F1 metric for the BFRC-PT corpus was the combination between the BR method and the RF algorithm. The combination between the BR method and the LSTM algorithm enabled the sixth best result among the 20 methods tested.

The different results obtained for the different corpus used demonstrate how the characteristics of each dataset influence the choice of the method and the most appropriate algorithm. For this reason, as future work we plan to extend experiments with more corpora, using other languages and labels. We also plan to use Deep Learning techniques combined with other methods of problem transformation, especially with techniques that use a class ensemble, such as RA*k*EL.

Acknowledgment

We would like to thank Siemens Ltd and NVIDIA Corporation for partially support this work.

References

[1] M.-Y. Day and Y.-D. Lin, "Deep learning for sentiment analysis on google play consumer review," in *IEEE Int. Conf. Inf. Reuse Integr.* IEEE, 2017, pp. 382–388.

[2] Y. Wang, M. Huang, L. Zhao *et al.*, "Attention-based lstm for aspect-level sentiment classification," in *Proc. Conf. Emp. Methods Natural Lang. Process.*, 2016, pp. 606–615.

[3] P. Desmet, "Measuring emotion: Development and application of an instrument to measure emotional responses to products," in *Funology*. Springer, 2003, pp. 111–123.

[4] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014.

[5] S. M. Liu and J.-H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. with Appl.*, vol. 42, no. 3, pp. 1083–1093, 2015.

[6] K. Song, S. Feng, W. Gao, D. Wang, L. Chen, and C. Zhang, "Build emotion lexicon from microblogs by combining effects of seed words and emoticons in a heterogeneous graph," in *Proc. ACM Conf. Hypertext & Social Media*. ACM, 2015, pp. 283–292.

[7] D.-A. Phan, H. Shindo, and Y. Matsumoto, "Multiple emotions detection in conversation transcripts," *Proc. Pacific Asia Conf. Language, Inform. Comput.*, p. 85, 2016.

[8] L. Wang, F. Ren, and D. Miao, "Multi-label emotion recognition of weblog sentence based on bayesian networks," *IEEJ Trans. Elect. Electron. Eng.*, vol. 11, no. 2, pp. 178–184, 2016.

[9] C. Pool and M. Nissim, "Distant supervision for emotion detection using facebook reactions," in *Proc. Workshop Comput. Modeling of Peoples Opinions, Personality, and Emotions in Social Media*. ACL, 2016, pp. 30–39.

[10] P. K. Bhowmick, A. Basu, P. Mitra, and A. Prasad, "Multi-label text classification approach for sentence level news emotion analysis." in *Proc. Int. Conf. Pattern Recognition and Mach. Intell.* Springer, 2009, pp. 261–266.

[11] Y. Zhang, L. Su, Z. Yang, X. Zhao, and X. Yuan, "Multi-label emotion tagging for online news by supervised topic model," in *Proc. Asia-Pacific Web Conf.* Springer, 2015, pp. 67–79.

[12] J. M. van der Zwaan, I. Leemans, E. Kuijpers, and I. Maks, "Heem, a complex model for mining emotions in historical text," in *Proc. Int. Conf. e-Science*. IEEE, 2015, pp. 22–30.

[13] L. Duan, S. Oyama, H. Sato, and M. Kurihara, "Separate or joint? estimation of multiple labels from crowdsourced annotations," *Expert Syst. with Appl.*, vol. 41, no. 13, pp. 5723–5732, 2014.

[14] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[15] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview," *Front. Comput. Sci.*, pp. 1–12, 2008.

[16] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, p. 333, 2011.

[17] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Eur. Conf. Mach. Learn.* Springer, 2007, pp. 406–417.

[18] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. Workshop Mining Multidimensional Data*, 2008, pp. 30–44.

[19] J. Fürnkranz, E. Hüllermeier, E. L. Mencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Mach. Learn.*, vol. 73, no. 2, pp. 133–153, 2008.

[20] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Proc. IEEE Int. Conf. Granular Computing*, vol. 2. IEEE, 2005, pp. 718–721.

[21] ——, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] M. M. Dosciatti, L. P. C. Ferreira, and E. C. Paraiso, "Anotando um corpus de notícias para a análise de sentimentos: um relato de experiência (annotating a corpus of news for sentiment analysis: An experience report)," in *Proc. Brazilian Symp. Inform. Human Language Technol.*, 2015, pp. 121–130.

[24] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[25] N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Silva, and S. Aluísio, "Portuguese word embeddings: Evaluating on word analogies and natural language tasks," in *Proc. Brazilian Symp. Inform. Human Language Technol.*, 2017, pp. 122–131.