Assessing Regression-Based Sentiment Analysis Techniques in Financial Texts

Taynan Maier Ferreira[1], **Francisco Caio Lima Paiva**[1], **Roberto Fray da Silva**[1], **Angel Felipe Magnossão de Paula**[1], **Anna Helena Reali Costa**[1], **Carlos Eduardo Cugnasca**[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais
Escola Politécnica da Universidade de São Paulo (USP)
Av. Prof. Luciano Gualberto, tv 3, 158 - Butantã - São Paulo, SP

{taynan.ferreira, francisco.paiva}@usp.br

{roberto.fray.silva, angel.magnossao}@usp.br

{anna.reali, carlos.cugnasca}@usp.br

Abstract. *Sentiment analysis (SA) is increasing its importance due to the enormous amount of opinionated textual data available today. Most of the researches have investigated different models, feature representation and hyperparameters in SA classification tasks. However, few studies were conducted to evaluate the impact of these features on regression SA tasks. In this paper, we conduct such assessment on a financial domain data set by investigating different feature representations and hyperparameters in two important models – Support Vector Regression (SVR) and Convolution Neural Networks (CNN). We conclude presenting the most relevant feature representations and hyperparameters and how they impact outcomes on a regression SA task.*

1. Introduction

The rise of the importance of social media and online content (e.g., Social Networks, Blogs, Online News) in modern society has led to the generation of an enormous amount of opinionated textual data and consequently increase in the number of sentiment analysis (SA) researches [Liu 2012]. Many domains have benefited from this expansion such as political [Wang et al. 2012], entertainment [Pang and Lee 2004], and others [Kim 2014, Zhang et al. 2018]. In our work, we will focus on the financial domain, which has been able to bring to light some important results and conclusions not only to its specific domain but also to the overall SA field [Bollen et al. 2011, Khadjeh Nassirtoussi et al. 2015].

Regarding SA tasks on the financial domain, previous work identified important issues such as the lack of standardized datasets and the predominance of classification tasks over regression [Khadjeh Nassirtoussi et al. 2014]. SemEval 2017 Task 5 addressed these problems and developed a financial domain dataset, composed of sentences from financial news headlines, for a SA challenge [Cortis et al. 2017, Davis et al. 2016]. Moreover, the developed dataset was labeled by financial experts, generating a gold standard sentiment score regression dataset. Nevertheless, because of the limitations of using experts for labeling data, this dataset is rather small, with a little more than a thousand

instances. On the other hand, most researches [Socher et al. 2013, Hu and Liu 2004, Yadollahi et al. 2017] are based on SA datasets with many thousands of instances.

There are three major classes of SA approaches [Medhat et al. 2014]: lexicon-based, machine learning based, and hybrid. The lexicon-based approach uses a sentiment lexicon (i.e., dictionary of word with associated prior sentiment knowledge) as a basis for defining rules that help to decide on the overall sentiment of a document [Bollen et al. 2011]. For example, considering SA classification tasks (in this case movie and product reviews), previous work achieved state-of-the-art results by creating a manually labeled general language domain (i.e., without highly technical language) sentiment lexicon called VADER and then using it together with some designer rules [Hutto and Gilbert 2014]. Other works went further and took into account the issue of the usage of general domain lexicons for domain-specific tasks [Khadjeh Nassirtoussi et al. 2014, Ruder et al. 2016, Hamilton et al. 2016] and generated the Loughran-Mcdonald (LM) lexicon, which is domain-specific (i.e., contains highly technical language) [Loughran and Mcdonald 2011].

In the case of the machine learning based approaches, linguistics features are extracted and used in statistical methods that can predict the sentiment of documents. Considering this approach, we define our choices of models, and feature representation techniques on some tendencies observed in SA systematic reviews work [Khadjeh Nassirtoussi et al. 2014, Zhang et al. 2018]. First, there is a predominance of the usage of support vector machine (SVM) models, which mainly uses the feature representation of bag-of-words (BOW) combined with TF-IDF [Khadjeh Nassirtoussi et al. 2015]. Second, there is a tendency for the employment of deep artificial neural network models such as the Convolutional Neural Network (CNN) [Kim 2014, Zhang and Wallace 2017]. Finally, there is increasing use of word embedding – such as GloVe [Pennington et al. 2014] and word2vec [Mikolov et al. 2013] - for SA tasks.

The hybrid approach combines the previous approaches by the incorporation of prior sentiment knowledge from lexicons into the feature representation space of machine learning models. Some systematic literature reviews identified few works use the hybrid approach [Medhat et al. 2014, Khadjeh Nassirtoussi et al. 2014]. Recent hybrid approach work investigated one technique for incorporating general domain lexicon information into a recurrent neural network (RNN) model for domain-specific tasks [Ruder et al. 2016]. However, to the best of our knowledge, no work systematically compared varied techniques for the incorporation of lexicon approach information. Moreover, no work thoroughly examined the differences between machine learning and hybrid approaches, especially for the regression task.

In the present work, we perform a thorough empirical examination of some of the feature representation, lexicons and machine learning models used both on the overall sentiment analysis field and on SemEval 2017 Task 5. We compare the machine learning approach with the hybrid one, using different sizes of the GloVe word embedding [Pennington et al. 2014], VADER non-specific sentiment lexicon [Hutto and Gilbert 2014], financial domain lexicon Loughran-Mcdonald (LM) [Loughran and Mcdonald 2011], and various hyperparameter values. The machine learning models used are SVR (i.e., regression adaptation of the SVM) and CNN. Our work

takes inspiration from similar previous work [Zhang and Wallace 2017] to address this issue of insufficient empirical work to guide practitioners' decisions.

Our results show evidence that, at least for this data set and lexicon information incorporation technique, there is no significant advantage in using a domain-specific lexicon over a general domain sentiment lexicon. Also, although there is an improvement of the hybrid approach over the machine learning one, the impact of choices such as word embeddings dimensions or hyperparameters are much more substantial.

2. Related Work

Previous work conducted an extensive evaluation of the impact of the word embedding dimension on a movie reviews classification SA task [Melamud et al. 2016]. It used a logistic regression classifier and the word2vec [Mikolov et al. 2013] word embedding to conclude that its dimension is an aspect of feature representation with significant impact on performance. However, machine learning models that achieved state-of-the-art on this same movie reviews task, such as CNN [Kim 2014], were not considered. In contrast, some prior empirical work compared the impact of using a CNN with different hyperparameters and word embeddings (GloVe and word2vec) for general language domain classification tasks [Zhang and Wallace 2017]. However, this work missed some important feature representation aspects, such as the mentioned word embedding size influence [Zhang and Wallace 2017]. So, we believe these works left a gap for empirically evaluating the impact of diverse feature representation techniques in a state-of-the-art machine learning model.

We will address this gap by examining how feature representation aspects – such as the word embedding dimension size and lexicon information – impact on state-of-the-art machine learning models (CNN and SVR). Our work also differs from these previous ones by tackling a regression SA task on the financial domain (The SemEval 2017 Task 5 [Cortis et al. 2017, Davis et al. 2016]). To the best of our knowledge, we are the first to perform such rigorous sensitivity analysis of state-of-the-art techniques, from multiple SA approaches (Machine Learning and Hybrid), on a regression domain-specific SA problem.

Regarding sentiment analysis hybrid approach, previous work combined the GloVe word embeddings with the VADER lexicon [Hutto and Gilbert 2014] information into a CNN model [Mansar et al. 2017]. However, this previous work didn't experiment with the financial domain Loughran-Mcdonald (LM) lexicon [Loughran and Mcdonald 2011]. Unlike this previous work, both the general domain and financial domain lexicons will be here employed.

Concerning the machine learning approach, previous work examined the impact of various word embedding-based features in SA classification tasks [Petrolito and Dell'Orletta 2018]. Aspects analyzed included how different embedding training corpus size, embedding training corpus domains, methods to combine word embeddings, among others, impacted classification outcomes. However, this work did not experiment with SA regression tasks. Furthermore, this research did not explore how different embedding-features interact with each other, i.e., how results may be affected by varying multiple features simultaneously.

3. Methodology

The chosen data set was from SemEval2017 Task 5, Subtask 2. It contains labeled financial news headlines for a Sentiment Analysis regression problem. It is composed of 1,142 financial headlines for the train set and 491 for the test set. Each headline is labeled with sentiments on a continuous scale from -1 (bearish) to +1 (bullish), with 0 being neutral.

The methodology adopted for this paper was divided into two main steps, as follows.

Data Preparation: we applied a stratified 5-Fold split, to obtain 5 sentiment-balanced folds for cross-validation. To ensure comparability across models and reproducibility, all models had at their input the same 5 folds, which were obtained using a fixed seed. Several established NLP data preparation steps were employed: stop words removal using the NLTK package; punctuation removal; lowercasing; tokenization using space as separators; and replacement of the companies' name with the word "company". All implemented models shared the same data cleaning process.

Model implementation: we implemented a total of 1.944 CNNs and 1.512 SVRs. This number of models was obtained by combining all the different possible feature representation and hyperparameters displayed in Table 1. The chosen feature representation and hyperparameters were selected considering the state-of-the-art SA tasks [Zhang and Wallace 2017] and the winning architectures of the SemEval task related to the data set used [Cortis et al. 2017, Davis et al. 2016]. We used the "Wikipedia 2014 + Gigaword 5" pre-trained GloVe word embedding [Pennington et al. 2014]. When using lexical dictionaries, each word available in the dictionary was represented by an n-dimensional vector: 7-dimensional for the Loughran-McDonald dictionary and 4-dimensional for VADER. To represent the sentiment of each sentence as a single vector, we average the sentiment vector of each word. The CNN architecture was composed of 2 Convolutional Layers followed by a single Dense Layer. When using dictionaries, GloVe and Dictionary vectors were concatenated before the Dense Layer, as shown in Figure 1. In CNNs, we use MSE as the loss function, the "Adam" optimizer, the ReLU activation function in the hidden layers, and the tanh function in the output layer.

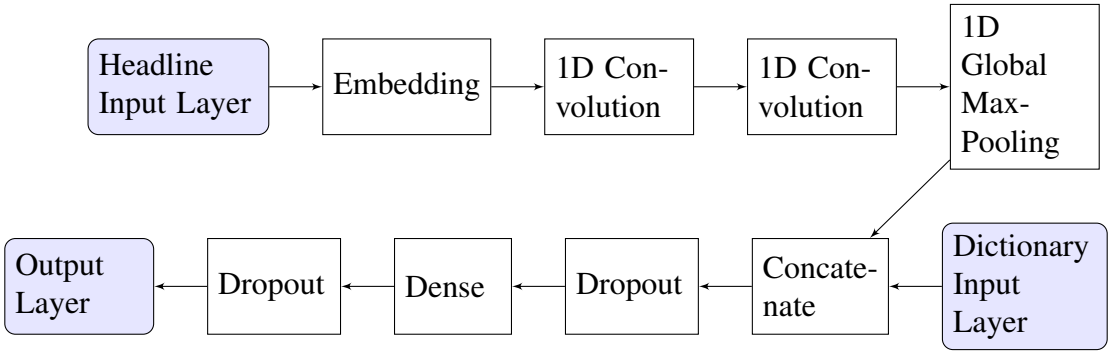

Figure 1. CNN architecture used in our experiments: 2 convolutional layers followed by single dense layer.

The SVR was implemented using the scikit-learn Machine Learning Library,

Table 1. Feature Representation and Hyperparameters evaluated on the different models

SVR
Input: TF-IDF and 50, 100, 200 and 300-dimensional GloVe word embeddings
Headline Representation: Sum and Mean of GloVe word embeddings
Dictionary: None, VADER and Loughran-McDonald (LM)
Kernel Type: rbf, sigmoid, poly
C (error penalty param): 0.01, 0.1, 1, 5, 10, 20, 50
Epsilon: 0.001, 0.01, 0.1
CNN
Input: 50, 100, 200 and 300-dimensional GloVe word embeddings
Dictionary: None, VADER and Loughran-McDonald (LM)
Batch size: 32, 64
Number of Filters: 128, 256, 384
Size of Filters: 2, 3, 4
Number of Neurons Dense Layer: 50, 100, 150
Dropout Rate: 0.3, 0.4, 0.5

whereas for the CNN we used Keras Deep Learning Library. All models were implemented in a server with the following technical specifications:

- CPU: Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz;
- GPU: GeForce GTX 1080 Ti;
- Memory: 4 x 16 GB DDR4 2133MHz.

4. Results and Discussion

In this section, we present and discuss the results obtained in our experiments. We will start presenting the effect of feature representation, followed by a discussion about hyperparameters. We close by discussing how the interaction of feature representation and hyperparameters impacts performance.

4.1. Effect of Feature Representation

In order to identify the global overall impact of feature representation across different models and setups, we condensed the different variations of the feature representation by taking the mean of the MSEs values over all experiments, grouping all the values according to the word embedding dimension size and dictionary utilized (e.g., first blue dot on the left side of the graphics of Figure 2 is the mean of the MSEs of all models using only a 50-dimensional GloVe embedding without dictionaries). Each line in the graphics represents a model setup configuration and each dot mark is a result of that configuration for that embedding size. The results are shown in Figure 2.

CNN results in Figure 2 show a clear trend of improvement of MSE mean values as the word embedding dimension is increased, with the biggest improvements happening from size 50 to 100 (roughly 7.5%) and from 100 to 200 (around 6.0%). The superiority of bigger word embeddings was observed not only by averaging the MSE along all hyperparameters, as in Figure 2: when ranking all CNN results by ascending MSE, the 4

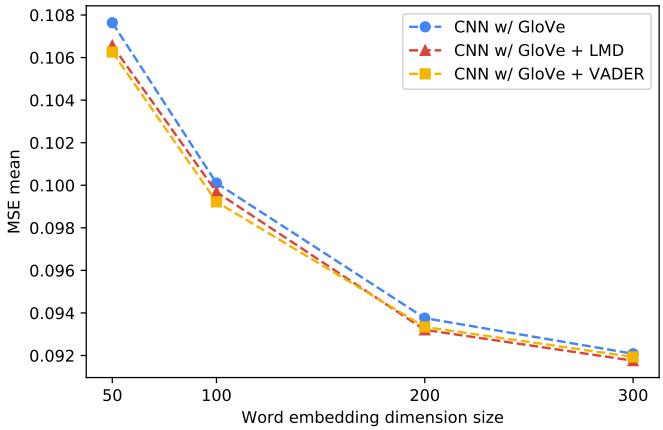

Figure 2. CNN mean MSE results varying word embedding dimension.

best results are CNNs with 300-dimensional word embeddings. Out of the 30 best overall results, only 6 are CNNs with 200-dimensional Word Vectors, the remaining being of 300 dimensions. Although the improvement from a word embedding dimension size of 200 to 300 is smaller (2%), it seems there could still be margin to improvement on larger embedding size, which may be explored in future work. This trend is in line with what other work found regarding the impact of word embedding dimension in CNNs trained for classification SA tasks [Melamud et al. 2016].

We observe a similar pattern of growing performance with an increasing embedding dimension on SVR results (Figure 3). Furthermore, the gap in performance observed between different embedding dimensions is still larger than that observed on the CNN: the use of 300-dimensional embedding almost halved the average MSE result observed on 50-dimensional ones.

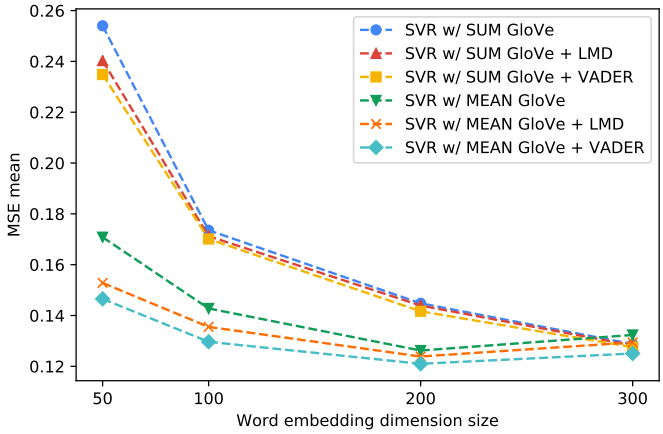

Figure 3. SVR mean MSE results varying embedding dimension.

Concerning the use of sentiment dictionaries, even though their use led consistently to better results than using only word embeddings, regardless of dimension sizes, the benefits of using dictionaries are marginal. We also observe that the benefits of using the dictionaries decrease as the embedding dimension sizes increase. Besides, it is important to notice that there seems to be no advantage in using a domain-specific dictio-

nary, as it is hypothesized in [Khadjeh Nassirtoussi et al. 2014]. The marginal benefit of dictionaries can be observed both on CNN and on SVR results.

Another important aspect to be taken into account, specifically for SVRs, since this model needs a fixed-size input, is how to combine the individual word embeddings to represent the whole sentence. As showed in Figure 3, we found strong evidence that this is a choice with a relevant impact on performance. Two of the most common ways to represent sentences from individual word embedding vectors is by taking the sum of the vectors or by taking their mean. Our experiments showed that taking the mean of the individual vectors resulted in generally superior outcomes. Nevertheless, we observe that this result is strongly dependent on the embedding dimension used: the difference between the two approaches decreases with increasing embedding dimension and completely disappears when using 300-dimensional embedding. The superiority of the mean representation is opposite to what other researchers found in their classification SA studies [Petrolito and Dell'Orletta 2018].

It is interesting to notice, however, that the aforementioned studies did not compare the sum and mean representations along with different embedding sizes, but only at a fixed dimension. This may be a major limitation, since, as we have shown, the gap between both approaches is strongly dependent on this factor. This finding suggests that the best word embedding aggregation method should be found considering the entire relevant hyperparameter and feature representation space, not only some arbitrary subset.

4.2. Effect of Hyperparameters

In this section, we present the impact of hyperparameters in model performance without taking into account its interactions with feature representation. We start by showing CNN results, followed by SVR.

4.2.1. Convolutional Neural Networks

As presented in Section 3, the combination of all hyperparameters and feature representation resulted in 1.944 different CNN models trained. In the following paragraphs, we exhibit the results of different hyperparameters configurations in CNNs.

Figure 4 shows the result of varying the number of filters along with the numbers of 128, 256 and 384. The results indicate a clear trend of better results as the number of filters increases.

Figure 5 displays the result obtained by varying the filter size in the CNN along with the values of 2, 3 and 4. In this scenario, smaller filter sizes brought better results.

Two different values of batch sizes were tested: 32 and 64. This hyperparameter had a significant impact on the outcomes, the smaller batch size leading to 8.5% superior results, on average.

As with other hyperparameters, the dropout rate was also varied and the chosen values were 0.3, 0.4 and 0.5. In contrast to the results of other hyperparameters, varying the dropout rate along the aforementioned range did not affect the result as much. The difference between the mean MSE of the best and worst dropout rate (0.4 and 0.5, respectively) was of only 1.2%.

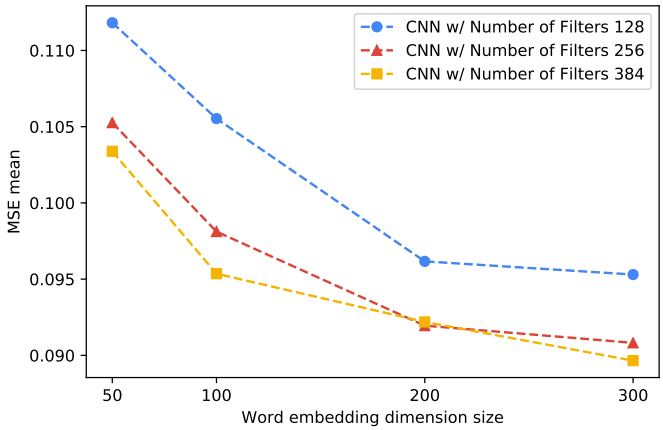

Figure 4. CNN MSE mean results of varying number of filters.

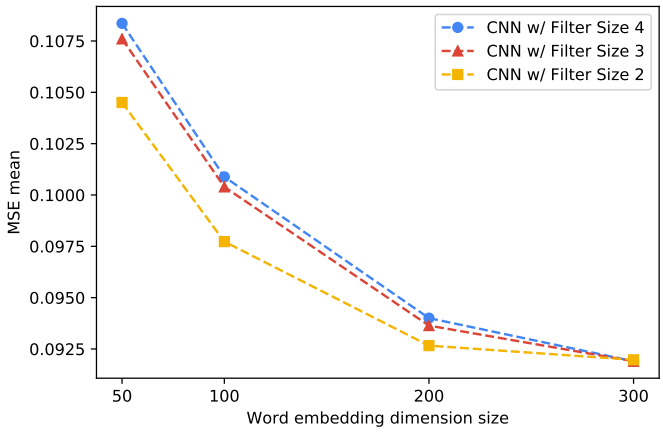

Figure 5. CNN MSE mean results of varying embedding size and filter size.

The number of neurons on the dense layers also proved to be a hyperparameter with irrelevant impact on the outcome of our experiments. Increasing this number yielded only immaterial superior results, on the order of 0.3%.

4.2.2. Support Vector Regressor (SVR)

In respect to the hyperparameters tested, the Kernel was the one with a greater impact on SVR results, the RBF Kernel leading to better average outcomes. The C penalty parameter was the second most important parameter in determining SVR performance.

We found no clear pattern that could lead to general conclusions or guidelines regarding do SVR hyperparameters, being grid search the best option to explore the hyperparameter space.

4.3. Interaction between Hyperparameters and Feature Representation

Besides studying the impact of varying individual hyperparameters and feature representation separately in CNN performance, we found useful investigating how the interaction of different hyperparameters and feature representation would affect results.

Figure 5 exhibit the impact of embedding dimension and filter size in CNNs outcomes. Here again, as expected, the bigger embedding dimension yielded better results. At least two other conclusions can be drawn from the results obtained. First, Figure 5 indicates that the embedding dimension has a greater impact in CNN performance when compared to filter size, since there is no overlap between the outcomes of different embedding sizes: the best result with 50 dimensions, for example, is still worse than the worst outcomes in any of the other greater dimensions. Another interesting pattern observed is the diminishing importance of filter size as the embedding dimension gets greater. Whereas the impact of the filter size is significant in the 50-dimensional embedding, this difference flattens out as the embedding dimension increases, getting irrelevant at the 300-dimensional embedding.

4.4. Comparison between best models

Table 2 shows the best result achieved by each of the Machine Learning models and feature representation when using their best-performing hyperparameters. The best overall performance was obtained by Model 6, a CNN with GloVe and the VADER Dictionary.

CNN achieved better outcomes than any SVR on 5-Fold average MSE results. SVR, on the other hand, was 10 to 20 times faster to train. Concerning feature representation, using word embedding performed better than using TF-IDF for SVRs, besides being also faster to train.

When comparing SVR and CNN best results, we see that their leading architectures' performance differs from 4.9%. This small difference could be at least partially attributed to the relatively small data set. Since CNNs are data-hungry techniques [Bilen and Vedaldi 2016], their advantage to more traditional Machine Learning techniques could fade away in such a data scarcity scenario.

Table 2. 5-Fold MSE, Test MSE and Mean Training Time for implemented models.

Model	**5-Fold MSE**	**Test MSE**	**Mean Training Time (s)**
Model 1: SVR with TF-IDF	0.095	0.1077	0.7531
Model 2: SVR with GloVe	0.0882	0.1019	0.1365
Model 3: SVR with GloVe and Vader	0.0883	0.1006	0.1118
Model 4: SVR with GloVe and LMD	0.0895	0.1013	0.1363
Model 5: CNN with GloVe	0.0862	**0.0942**	2.49
Model 6: CNN with GloVe and Vader	**0.0853**	0.0949	2.907
Model 7: CNN with GloVe and LMD	0.0855	0.0968	2.493

5. Conclusions and Future Work

To the best of our knowledge, our paper was the first one to systematically study and analyze hyperparameter sensitivity and feature representation in a SA regression problem. We could not find, after a thorough review of the literature, other researchers that proposed an evaluation of regression SA considering different models, preprocessing techniques and dictionaries.

We summarize our main findings and conclusions below:

Larger word embedding dimensions consistently led to better model performance. This finding was consistent regardless of the Machine Learning model used and the different hyperparameter options. [Melamud et al. 2016] came to similar conclusions in SA classification tasks.

Models that need a fixed-size input should use the mean of individual word vectors. When using word embedding together with classical Machine Learning models, such as SVR, which need a fixed-size input, one should consider the mean of the individual word vectors as the first choice of feature representation. Representing sentences as the mean of the individual word embeddings led to superior results in a great variety of hyperparameter choices, though its advantage diminished on larger embedding dimensions.

Lexical dictionaries bring an only slight increase in performance The use of lexical dictionaries resulted in better model performance in all techniques, although the extra gain was generally small compared to other aspects, especially those related to the word embeddings.

The best CNN models led to only slightly better results compared to the best SVR model. This is probably due to the relatively small data set combined with the fact that CNNs are data-hungry techniques [Bilen and Vedaldi 2016]. Therefore, when dealing with relatively small data sets, considering the time-consuming effort of adjusting all relevant hyperparameters in a CNN, one should consider using a classic Machine Learning model such as SVR. This is especially true in real scenarios, where training time and time for financial forecasting can be scarce.

Our contribution to the state-of-the-art expands the knowledge of sentiment analysis by complementing its pre-existing classification research with an unprecedented systematic regression investigation on the SA task. Furthermore, our study proved the importance of doing such research exploring all the relevant hyperparameter space, not only within some arbitrary subspace, at the risk of taking local phenomena as general truths.

For future work, we intend to broaden our efforts by advancing in several different aspects. First, we want to confirm whether the conclusions here drawn are resilient and also hold on other data sets. Second, we would like to expand both the set of varied hyperparameters and the range in which they are varied in this study. Moreover, we would like to try sentiment-aware word embeddings and to find what aspects, if any, contrasts to the results found in the literature for SA classification tasks.

Acknowledgments

This work is being carried out with the support of *Itaú Unibanco S.A.*, through the scholarship program of *Programa de Bolsas Itaú* (PBI), linked to the *Centro de Ciência de Dados* (C^2D) of *Escola Politecnica da USP*. We also would like to thank CNPq (Proc. No. 425860 / 2016-7 and N. 307027 / 2017-1) for the support.

References

Bilen, H. and Vedaldi, A. (2016). Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., and Davis, B. (2017). SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Stroudsburg, PA, USA. Association for Computational Linguistics.

Davis, B., Cortis, K., Vasiliu, L., Koumpis, A., Mcdermott, R., and Handschuh, S. (2016). Social Sentiment Indices Powered by X-Scores. In *ALLDATA 2016, The Second International Conference on Big Data, Small Data, Linked Data and Open Data*, Lisbon, Portugal.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Hutto, C. J. and Gilbert, E. (2014). VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth international AAAI conference on weblogs and social media*, pages 216–225.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1):306–324.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability ? Textual Analysis , Dictionaries , and 10-Ks. *Journal of Finance*, 66(1):35–65.

Mansar, Y., Gatti, L., Ferradans, S., Guerini, M., Staiano, J., Solutions, F. F., and Kessler, F. B. (2017). Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–6.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Melamud, O., McClosky, D., Patwardhan, S., and Bansal, M. (2016). The role of context types and dimensionality in learning word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Lin-*

guistics: Human Language Technologies, pages 1030–1040, San Diego, California. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe : Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Petrolito, R. and Dell'Orletta, F. (2018). Word embeddings in sentiment analysis. In *CLiC-it*.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea. Association for Computational Linguistics.

Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.*, 50(2):25:1–25:33.

Zhang, L., Wang, S., and Liu, B. (2018). Deep Learning for Sentiment Analysis : A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):1–25.

Zhang, Y. and Wallace, B. C. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 253–263.