Combining Representations For Effective Citation Classification

Claudio Moisés Valiense de Andrade
Federal University of Minas Gerais
claudio.valiense@dcc.ufmg.br

Marcos André Gonçalves
Federal University of Minas Gerais
mgoncalv@dcc.ufmg.br

Abstract

We describe our participation in two tasks organized by WOSP 2020, consisting of classifying the context of a citation (e.g., background, motivational, extension) and whether a citation is influential in the work (or not). Classifying the context of an article citation or its influence/importance in an automated way presents a challenge for machine learning algorithms due to the shortage of information and inherently ambiguity of the task. Its solution, on the other hand, may allow enhanced bibliometric studies. Several text representations have already been proposed in the literature, but their combination has been underexploited in the two tasks described above. Our solution relies exactly on combining different, potentially complementary, text representations in order to enhance the final obtained results. We evaluate the combination of various strategies for text representation, achieving the best results with a combination of TF-IDF (capturing statistical information), LDA (capturing topical information) and Glove word embeddings (capturing contextual information) for the task of classifying the context of the citation. Our solution ranked **first** in the task of classifying the citation context and *third* in classifying its influence.

1 Introduction

Data science is becoming more and more popular with its largest data community being *Kaggle*[1], a platform that hosts several data mining and machine learning tasks and challenges. In 2020, the 8th International Workshop on Mining Scientific Publication (WOSP)[2], through Kaggle, promoted two challenges consisting of: 1) classifying the context of a citation in one of the six existing classes (e.g., background, motivational, extension) and 2) a binary task where the goal is to identify the importance of a citation for a given work. The competition overview was presented by Kunnath et al. (2020) (N. Kunnath, 2020).

An example of the citation context taken from the dataset is:*"In the future we are planning to experiment with different ways of calculating relatedness of the sequences to the descriptions, such as with computing similarity of embeddings created from the text fragments using approaches like Doc2Vec (#AUTHOR_TAG and Mikolov, 2014)"*, where #AUTHOR_TAG tag means the quote being classified. In this case, for the context classification challenge, this citation belongs to the Future class, and for the influence classification task, this citation is considered influential.

For the sake of the defined classification tasks, the citation text can be represented in several ways (e.g. TF-IDF, word embeddings, text graph), each representation capturing or focusing on a different aspect of the task. For instance, the traditional TF-IDF representation captures statistical aspects of the text and the specificity of certain words in the collection (IDF component). Topic modeling strategies such as LDA identify patterns of recurrent topics (i.e., clusters of words) in the text. Word embeddings are vectorial representations of words, sentences and whole documents aimed at capturing word co-occurrence patterns and contextual information.

Our main hypothesis here is that these different sources of information are somewhat complementary and, when combined, have the potential to enhance classification effectiveness. By exploring such ideas, we were able to reached the **first** place in the multiclass classification task promoted by *WOSP* and *third* in the binary influence classification task, with further improvements after the closing of the challenge, as we shall see.

In the competition there are two types of scores:

[1] http://www.kaggle.com
[2] https://wosp.core.ac.uk/jcdl2020

1) the public score, calculated based on 50% of the test data and 2) the private score, based on the results of the other 50% and only displayed after the closing of the challenge. Using a combination of TF-IDF, LDA and Glove embedding representations, as aforementioned, along with a Passive-Aggressive classifier (Crammer et al., 2006), our final result improved by 3.62% the scores of the best isolated representation in the public score, while for the private score this difference was up 6.91%. For the task of assessing the importance of the citation, we obtained an improvement of 1.39% for the public score and 3.07% for the private one, by adding a feature that identifies that the author of the quote is the same author who is making the citation.

To guarantee the reproducibility of our solution, all the code is available on github[3][4] and we create an image through the docker with the entire process configured.

2 Related Work

Some papers addressed the problem of classifying the citation for context and influence. Jurgens et al. (2018) (Jurgens et al., 2018) aimed at classifying the context of a citation, similarly to the challenge, into six possible classes. To train the model, they used structural information from the text, lexical, grammatical, etc. In that work, specific terms have higher weights in relation to the others, for example, "we extend" have more importance in the classification of the context "Extends". Unlike their work, we exploit only information from titles and the context of the citation with combinations of representations.

In Valenzuela et al. (2015) (Valenzuela et al., 2015), if the citation is connected to related works or is used for comparison between methods, the author considers it as an incidental citation. If the cited method is used in the work or the current work is an extension of the cited one, the citation is considered important. These are the two possibilities in the challenge of the influential classification competition. In that work, the authors used 12 features, among them, the citation count per section, similarity between abstracts, etc.

Pride et al. (2017) (Pride and Knoth, 2017) also dealt with the binary task of classifying an influential citation. In that work, the authors expose the problem of extracting data from *pdf* when there is no structured data. Their work analyzed the features of previous works and the impacts of adding specific features to the model. Unlike Valenzuela and Pride's works, we do not use any information other than the titles and context of the citation, data provided through the Kaggle.

3 Methodology

In this section we present our methodology, consisting of of applying preprocessing to the data, creating different representations to explore the importance of individual words (TF-IDF), group similar concepts shared by terms (LDA) and explore the semantics and context of terms (Glove). After this step, we join (concatenate) the representations to train the classifier, aiming to predict the class of the test set.

The dataset contains eight fields, for each document we combine fields Citing Paper Title, Cited Paper Title and Citation Context (hereafter called *citation text* or simply citation (Cit)) on a single line, separated by space. We use this file as input to the data preprocessing algorithm, which consists of: 1) turning the text into lowercase, 2) removing accents from words and 3) applying a parser that replaces specific strings with tags fixes, for example, number by "parserNumber".

After preprocessing the citation text, we initially use the TF-IDF (Luhn, 1957) representation. It exploits both, unigrams and bigrams. where the bigram is the combination of the current term and the next one. In more details, TF quantifies the frequency of the terms (unigrams or bigrams) and the IDF measures their inverse frequency of document, giving higher weights to terms that occur less frequently in documents (higher discriminative power).

Another representation we use is the Latent Dirichlet Allocation (LDA) with online variational Bayes algorithm (LDA) (Hoffman et al., 2010). It groups terms into similar concepts called topics. Topics are learned as a probability distribution on the words that occur in each document, using as input the original TF-IDF representation. For each document, a score is associated with each topic and a citations may be seen as a combination of topics with different likelihoods.

We tested some values for the hyperparameter that defines the number of topics in the dataset.

[3] https://github.com/claudiovaliense/wosp_2020_3c-shared-task-purpose

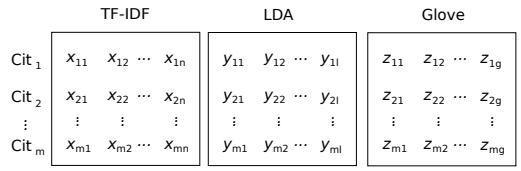
[4] https://github.com/claudiovaliense/wosp_2020_3c-shared-task-influence

We tested 6, 10, 50 and 100 topics. The one that produced the best results in the training data was 6 topics. That is, the algorithm creates 6 groups of words in the training and at the moment it receives a new test citation Cit_i, it assigns a score (likelihood) to each of the groups for Cit_i.

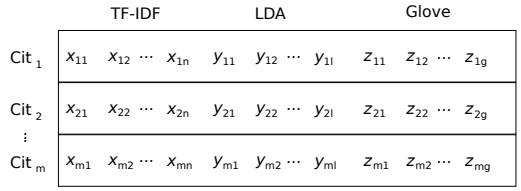
The last exploited representation was Glove embeddings (Pennington et al., 2014), an unsupervised learning algorithm to obtain vector representations for words. For each term *t* present in a citation *cit*, we average (pool) their respective Glove vectors to create a representation for the whole citation.

Each textual representation generates a number of features for each citation. In TF-IDF, this number corresponds to the amount of unique terms existent in the entire dataset. For LDA, this is the number of defined topics, in our case 6. For the Glove embeddings we use the vector representation with 300 dimensions, thus generating 300 features for each citation. Figure 1a presents the features for each citation cit_i, where m represents the number of citations and n, l and g the number of features generated by each representation.

Finally, to combine the representations, we use a simple concatenation of all available representations (Feature Union[5]) resulting in a single vector, as shown in Figure 1b.



(a) Representations.



(b) Feature Union.

Figure 1: Combining Representations.

4 Experiment

The dataset was created based on the methodology developed by Pride et al. (2019) (Pride et al., 2019). By means of an automatic system, authors (or external evaluators) can select to which group the citations are most related to (context) and whether the citation is described as central to the work.

Based on this methodology, 3000 citations with defined classes were made available through the Kaggle for training along with 1000 test citations that should be classified according to the defined classes. Table 1 describes the training data, with: Number of citations (|D|), Median of the amount of term (M(T)), Number of classes (|C|), Number of citations of the largest (Max(C)) and smallest (Min(C)) class. Note that the dataset is very unbalanced – the largest class has 1648 citations while the smallest contains only 62.

Table 1: Dataset Metadata.

Name	|D|	M(T)	|C|	Max(C)	Min(C)
Context	3000	55.00	6	1648	62
Influence	3000	55.00	2	1568	1432

4.1 Classification and Parameter Tuning

Among the classifiers we tested, Passive Aggressive, Stochastic Gradient Descent (SGD) and Linear SVM presented the best results in preliminary experiments and were selected to be used in the challenge. For each of them we optimized the hyperparameter through stratified cross-validation (10 folds) in the training set. For the Passive-Aggressive and Linear SVM classifiers, we evaluated the *C* parameter varying among $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}]$ and for SGD we vary the *alpha* parameter among $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$. We used a potency of 10 to avoid overfitting the classifier.

Table 2 shows the result of the process of tuning the parameters with the cross-validation procedure. We present the parameters that obtained the highest scores in cross-validation, the macro F1 score, in parentheses, the standard deviation and the time in seconds spent by each classifier. As can be seen, Linear SVM is about 5 times slower than Passive-Aggressive, while there is a statistical tie in the final (training) result. Since we would need to test many alternatives and configurations in our trials, we decided to choose Passive-Aggressive as the final classifier for the challenge. The Passive-Aggressive algorithms are a family of algorithms for large-scale learning not requiring a learning rate (Crammer et al., 2006). However, contrary to the Perceptron, they include a regularization parameter C.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html

Table 2: Result of classifiers in the evaluation.

Classifier	Best Parameter	Macro-F1	Time (s)
PA	C [10^0]	0.1846 (0.044)	94.36
SGD	alpha [10^{-4}]	0.1740 (0.033)	21.38
SVM	C [10^0]	0.1953 (0.049)	486.75

Table 3: Kaggle Score

Method	Public	Private	Task
tfidf	0.19829	0.19425	Context
lda	0.12923	0.15826	Context
glove	0.12047	0.11489	Context
tfidf+lda	0.19124	0.19572	Context
tfidf+glove	0.19945	0.20037	Context
tfidf+lda+glove	**0.20548**	**0.20560**	**Context**
tfidf	0.59108	0.54747	Influence
lda	0.30458	0.32249	Influence
glove	0.30458	0.32249	Influence
tfidf+lda	0.32707	0.36156	Influence
tfidf+glove	0.30458	0.32249	Influence
tfidf+lda+glove	0.30458	0.32249	Influence
tfidf+same_author	**0.59932**	**0.56431**	**Influence**

4.2 Result

In the Kaggle challenge, the evaluation was based on Macro-F1, probably due to the high skewness of the Context task. For each new submission the score is calculated based on 50% of the test data (public score). The results of the other 50% (private score) were only displayed at the closing of the challenge. The final result of the competition was based on the private score. Table 3 presents the results of the individual representations as well as their combinations, considering the public and private scores.

For the classification of topics, the strategy that presented the best results used the combination of the three aforementioned representations – TF-IDF, LDA and GLOVE. Notice that in this task, the performance of TF-IDF is already high, better than LDA and Glove.

In the classification of influential citations, TF-IDF alone produces the best results. Combinations of representations using LDA, Glove or both, showed a reduction in the final score. In this task, the effectiveness of both LDA and Glove are far from TF-IDF, about 50% less effective. We hypothesize that the concatenation of the representations produce a very high dimensional space that, along with the not so good performance of LDA and Glove, exacerbates issues of noise and overfitting in this binary task. We will further investigate this in future work. After the submission deadline, we added a feature that captured whether the author being cited is the same author of the article that quotes, this feature has improved the final result (tfidf+same_author).

We should stress that the excellent performance of TF-IDF alone is consistent with recent results that show that TF-IDF, when coupled with a strong, properly tuned classifier, is still one of the best text representations, better than certain word embeddings for classification tasks, (Cunha et al., 2020).

5 Conclusion

In this paper we described our participation in the citation classification tasks organized by WOSP 2020. We focused on evaluating combinations of textual representations – statistical information with TF-IDF, topical with LDA, and contextual and co-occurence information with Glove word embeddings – and the impact of each one on the final result. Our solution relied on exploring multiple, potentially complementary, representations to add their benefits as they potentially capture different textual aspects. We use the Passive-aggressive classifier, the best and faster in a preliminary evaluation for the task, optimizing its hyperparameters through stratified folded cross validation within the training set. TF-IDF demonstrated to be a very powerful representation when used with a strong, properly tuned classifier, confirming recent results that it may better than certain alternatives (e.g., embeddings) for specific tasks (Cunha et al., 2020). But its combination with other representations indeed did help to improve results, as initially hypothesized. Overall, our solution achieved very good results, reaching the **first** place in the task of classifying the context of a citation and *third* in the classification of influential citations (with post-deadline improvements).

As future work, we intend to evaluate combinations with new representations, e.g., MetaFeatures (Canuto et al., 2018; Canuto et al., 2016, 2019) and Cluwords (Viegas et al., 2018, 2019, 2020). Due to the shortage of information, enhancing citation data with automatic tagging information (Belém et al., 2019, 2014, 2011) seems as a promising strategy to obtain more data.

Acknowledgments

This work was partially supported by CNPq, Capes and Fapemig.

References

Fabiano Belém, Eder Ferreira Martins, Tatiana Pontes, Jussara M. Almeida, and Marcos André Gonçalves. 2011. Associative tag recommendation exploiting multiple textual features. In *Proceeding of the 34th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 1033–1042.

Fabiano Muniz Belém, André G. Heringer, Jussara M. Almeida, and Marcos André Gonçalves. 2019. Exploiting syntactic and neighbourhood attributes to address cold start in tag recommendation. *Inf. Process. Manag.*, 56(3):771–790.

Fabiano Muniz Belém, Eder Ferreira Martins, Jussara M. Almeida, and Marcos André Gonçalves. 2014. Personalized and object-centered tag recommendation methods for web 2.0 applications. *Inf. Process. Manag.*, 50(4):524–553.

S. Canuto, D. X. Sousa, M. A. Gonçalves, and T. C. Rosa. 2018. A thorough evaluation of distance-based meta-features for automated text classification. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2242–2256.

Sérgio D. Canuto, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *Proc. of the 9th ACM Conf. on Web Search and Data Mining*, pages 53–62.

Sérgio D. Canuto, Thiago Salles, Thierson Couto Rosa, and Marcos André Gonçalves. 2019. Similarity-based synthetic document representations for meta-feature generation in text classification. In *Proc. of the 42nd ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR 2019*, pages 355–364.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585.

Washington Cunha, Sérgio Canuto, Felipe Viegas, Thiago Salles, Christian Gomes, Vitor Mangaravite, Elaine Resende, Thierson Rosa, Marcos André Gonçalves, and Leonardo Rocha. 2020. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management (IP&M)*, 57(4).

Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. In *Proc. of the 23rd Conf. on Neural Information Processing Systems - Volume 1*, page 856–864.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

H. P. Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317.

David; Gyawali Bikash; Knoth Petr N. Kunnath, Suchetha; Pride. 2020. Overview of the 2020 wosp 3c citation context classificationtask. in: Proceedings of the 8th international workshop on mining scientific publications. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

David Pride, Petr Knoth, and Jozef Harag. 2019. ACT: an annotation platform for citation typing at scale. In *19th ACM/IEEE Joint Conf. on Digital Libraries, JCDL 2019*, pages 329–330. IEEE.

David T. Pride and Petr Knoth. 2017. Incidental or influential? - a decade of using text-mining for citation function classification. In *ISSI*.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*, volume WS-15-13 of *AAAI Workshops*.

Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proc. of the 12th ACM Conf. on Web Search and Data Mining*, page 753–761.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo C. da Rocha, and Marcos André Gonçalves. 2020. Cluhtm - semantic hierarchical topic modeling based on cluwords. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online*, pages 8138–8150.

Felipe Viegas, Washington Luiz, Christian Gomes, Amir Khatibi, Sérgio D. Canuto, Fernando Mourão, Thiago Salles, Leonardo C. da Rocha, and Marcos André Gonçalves. 2018. Semantically-enhanced topic modeling. In *Proc. of the 27th ACM Conf. on Information and Knowledge Management, CIKM 2018*, pages 893–902.