Extended Pre-Processing Pipeline For Text Classification: On the Role of Meta-Features, Sparsification and Selective Sampling

Washington Cunha[1], **Leonardo Rocha**[2], **Marcos A. Gonçalves**[1]

[1] DCC - Universidade Federal de Minas Gerais (UFMG)
[2] DCOMP - Universidade Federal de São João del-Rei (UFSJ)

{washingtoncunha,mgoncalv}@dcc.ufmg.br, {lcrocha}@ufsj.edu.br

Abstract. *Pipelines for Text Classification are sequences of tasks needed to be performed to classify documents. The pre-processing phase of these pipelines involves different ways of manipulating documents for the learning phase. This* ***Master Thesis*** *introduces three new steps into the traditional pre-processing phase: 1) Meta-Features Generation; 2) Sparsification; and 3) Selective Sampling. Our experimental results, based on more than 5.600 measurements, show that our proposal can achieve significant gains in effectiveness when compared to the traditional TF-IDF representation (up to 52%) and word embeddings (up to 46%), at a much lower cost (9.7x faster). Our Master Thesis also includes a thorough and rigorous evaluation of the trade-offs between cost and effectiveness associated with the introduction of these new steps into the pipeline, as well as a comprehensive comparative experimental evaluation of many alternatives. This thesis falls under the topics of (i) Document Management and Classification, (ii) Information Retrieval Models and Techniques, (iii) and Text Database of the SBBD Call for Papers.*

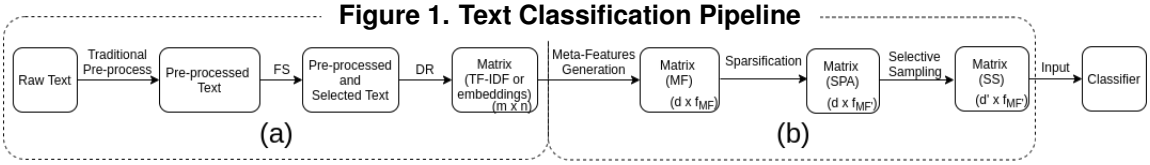
1. Introduction (Problem Statement)

Classification pipelines are sequences of steps needed to be performed for running a classification task. Briefly, ATC consists of mapping objects (e.g., textual documents, images) into a set of predefined categories. In this Master Thesis, we are specially concerned with *pre-processing steps* of *text classification* pipelines. **Figure 1(a)** illustrates a typical text classification pipeline. We usually apply some pre-processing to the raw textual content of documents, such as lower-casing, punctuation, stopwords removal, and stemming. Then, an additional feature selection step may be performed, keeping only the most informative words. After that, a $m \times n$ document-term matrix representation (or some latent term encoding) of the document is built. The most common representation of this matrix in text classification exploits the so-called TF-IDF paradigm. The major problems with the TF-IDF matrix representation have to do with its high dimensionality and sparseness. Other alternative representations that aim at producing a more compact space in terms of latent dimensions (e.g., distributional word embeddings) do exist [Mikolov et al. 2018]. However, their unambiguous contribution to text classification tasks has not yet been fully established, as comparisons are not made with standard benchmarks following rigorous scientific procedures[1]. In any case, the complexity of the representation is directly associated with the costs involved in the whole classification

[1]Issues regarding poor or inadequate experimental evaluation setups as well as improper or unfair baseline tuning have recently raised serious questions about the real value of complex deep learning solutions in areas related to information retrieval [Dacrema et al. 2019, Cunha et al. 2021].

process. This is particularly important given the development of recent techniques such as (word) embedding and deep neural networks, which require a considerable amount of data and computational power to properly work. These methods cause large increases in the cost of training, validation and actual classification of unknown (test) instances.

Figure 1. Text Classification Pipeline



It used to be the case, not in the distant past, that in text classification tasks the cost of training a model would not be a dominant factor in choosing a particular algorithm or representation, as the training process would be run just once and in a batch mode. This scenario has changed in the last few years, mainly due to the tremendous increase in the complexity of the representations (i.e. word embeddings), the complexity of the models (i.e. deep learning), the size of the datasets and the surge of new applications such real-time text stream classification [B.-Naranjo et al. 2021]. Although the possibility of exploiting cloud computing (e.g. AWS and GCP[2]) and the great computational advances (e.g. CPU and GPU architectures), these resources are still expensive, mainly for researchers at universities and small or medium-sized companies. Furthermore, suppose the accuracy is already sufficient for a given application. In that case, one may not need to run more complex models, which would require much more additional time or cost to train[3]. Finally, for some applications in which there are rapid changes in some assumptions of the model (e.g., class distributions, term distributions, etc) or in real-time stream-based applications in which recent events may affect such assumptions, a periodical retraining is required, which makes the training time a non-negligible factor. In sum, *there is a tradeoff between effectiveness and cost* and this has been acknowledged by recent work [Kastrati et al. 2019, Cunha et al. 2021]. Therefore, in our work, we consider the training time as an essential analysis factor and the gains in performance that can be obtained with the proposed pipeline without (or with minimal) losses in effectiveness.

Automatic Text Classification (ATC) models are considered the basis of several applications, such as information organization, document engineering, information extraction and representation, content recommendation and others. Therefore, we can state that the generation of the well-built classification models has a significant positive impact on the quality of all the tasks previously mentioned. In this Master Thesis, we are not trying to hide the complexity of configuring the classification pipeline. Instead, our concern is to manage issues related to the trade-off between classification effectiveness (e.g., high accuracy) and the costs involved to achieve it. This aligns with recent work [Schoenfeld et al. 2018] that shows that by properly working on the pre-processing phase of the pipeline, one can achieve significant gains in performance, even if effectiveness is not improved. Therefore, we investigate issues related to the *introduction, orchestration and combination* of **three new steps**[4](Meta-Feature Generation, Sparsification and Selective Sampling - Figure 1(b)) into the standard pre-processing phase of text classification pipelines to improve effectiveness while, at the same time, reducing associated costs.

[2]AWS: https://aws.amazon.com/pt/ec2/ and Google GCP: https://cloud.google.com/products/compute/

[3]https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

[4]The source code of each step is available on: https://gitlab.com/waashk/extended-pipeline

2. Contributions

The **first contribution** of this Master Thesis is the proposal of (distance-based) Meta-Feature (MF) generation as an *explicit* pre-processing step in a text classification pipeline. This step reduces the dimensionality and sparsity of the original TF-IDF document-term matrix by producing a potentially more informative denser space. Indeed, distance-based meta-features are very effective in text classification tasks [Canuto et al. 2018]. However, they have been proposed as an end in itself, not a way to achieve a more compact and richer representation in a text classification pipeline. As such, our contribution lies in re-thinking its role in a chain of data transformations before actual learning. By having this as an explicit goal, we introduce other contributions in this realm: (i) investigations regarding the dimensionality and density of the MF representation; (ii) definitive studies on the effectiveness-cost trade-offs associated with the introduction of this step into the pipeline. *Those aspects have not been previously studied.*

The **second novel contribution** of our Master Thesis is the introduction of a new *sparsification* step of the MF matrix. Although less dimensional and more informative , the MF representation is dense since all features have a non-zero value[5]. This may impact some classifiers' learning time, mainly when applied to datasets with many classes due to the increase in density. Moreover, the danger of incurring in the "curse of dimensionality"(making the classes less separable) in these denser spaces is high, mainly if many of the low-valued meta-features introduce more **noise**[6] [B.-Naranjo et al. 2021] than helpful information. For these reasons, we propose a procedure to transform such a denser representation into a sparser one while keeping the same dimensionality and potentially reducing noise. This aligns with recent work [Zamani et al. 2018] that aims at making dense latent embeddings representations sparser for the sake of performance in search tasks. The goal is to provide reductions in cost/time, with no significant losses in effectiveness.

Most of the previous related work (e.g., on feature selection) is column-oriented, but few are concerned with the lines (documents) of the matrix. Our **third novel contribution** is the introduction of a selective sampling (SS) step aimed at reducing the lines of the low-dimensional, denser matrices obtained in the previous MF step (with and without sparsification). For this, we propose to use the SOTA compression-based SS method (Cover) that has been originally proposed for the learning-to-rank task but has never been used in the ATC task. The main advantages of this method include: i) it is data-driven and independent of any learning framework; ii) it is computationally cheap for dense representations and scales well to large datasets; and iii) it can compress the original data in a much smaller representation without loss of information by selecting the "best" documents.

The **fourth contribution** is a thorough comparative study among many methods, representations and approaches along with several analyses regarding "how"s and "why"s. We provide a rich and complete set of experiments (**5.600 measurements**[7]) in order to answer the following research questions (**RQs**), associated with our three first contributions, which emphasize the impact of the orchestrated use of the proposed pipeline steps:

• ***RQ1*** *What are the gains in terms of the effectiveness of the MFs when compared to the original TF-IDF representation?*

[5]For more details, see Section 3.1 of derived publication [Cunha et al. 2020].

[6]Noise not only affects effectiveness but may also affect efficiency, regarding model parameterization, since the search for suitable parameters to cope with non-linearly separable cases can be higher.

[7]Experiments on 35 representations/classifiers combinations, four datasets, four metrics and ten folds.

• ***RQ2*** *What are the costs associated with this new step?*
• ***RQ3*** *How do the MFs compare with other alternative low-dimensional representations (e.g., word embeddings) in terms of effectiveness and incurred costs?*
• ***RQ4*** *Can the more compact (SS) and less dense (SPA) MF representations reduce costs without loss of effectiveness?*
• ***RQ5*** *Is the previous step MF (create a low-dimensional representation) really necessary to apply the proposed SS technique? Or could we use it with the original TF-IDF matrix?*
• ***RQ6*** *Can the compact representation induced by the proposed pipeline benefit different state-of-the-art (SOTA) classifiers?*

The results of this Master Thesis are summarized in a paper published in the *Information Processing and Management* **(IP&M)** *(h5-index:46, Imp. Fac:4.787, Qualis A1)* [Cunha et al. 2020], a worldwide leading journal in *Information Retrieval*. This Master Thesis also directly contributed in a second **IP&M** paper [Cunha et al. 2021], covering a comprehensive comparative study of cost-effectiveness of neural and non-neural approaches and representations for ATC. Our Master Thesis also contributes to other papers in important conferences, such as **WWW** [Cunha et al. 2018], **CIKM** [Viegas et al. 2018, Mendes et al. 2020], **WSDM** [Viegas et al. 2019] and **ACL** [Viegas et al. 2020].

3. Experimental Evaluation

3.1. Experimental Setup: We consider six textual datasets widely used in literature in the domains of sentiment analysis and topic classification: (i) **WebKB** , Web pages collected from Computer Science departments; (ii) 20NewsGroup (**20NG**) newsgroup documents, (iii) **ACM** Digital Library; (iv) **Reuters**, composed of news articles; (v) **SOGOU** news corpora containing 500K news articles in various topic channels; and (vi) **IMDB** reviews composed of $348,415$ user reviews about 50K movies. We evaluate the effectiveness of our proposal with Micro Averaged F1 and Macro Averaged F1. All experiments were executed using a 10-fold cross-validation procedure.[8] Parameters were set via cross-validation on the training set and the effectiveness of the algorithms running with distinct types of features was measured in the test partition. We adopted the LibLinear implementation of the SVM classifier, as it still is one of the best text classifiers capable of dealing with both high and low dimensional representations. As we saw in [Cunha et al. 2021], SVM was *superior to neural alternatives such as **BERT** and **XLNet** in the tested datasets* for reasons explained in [Cunha et al. 2021]. For feature selection (**FS**), we consider the importance score of Random Forests applied [Louppe et al. 2013]. We consider only this strategy since the authors argue that it could be considered the SOTA and for a question of scope delimitation of this work. Next, we have the cross-validation established parameters values for: (i) sparsification and (ii) selective sampling. To compare the average results on our cross-validation experiments, we assess the statistical significance employing a paired t-test with 95% confidence and Bonferroni correction to account for multiple tests. and the Friedman-Nemenyi-Test for multiple comparisons of mean rank sums.

3.2. Answers to Research Questions. Next, we will provide a summary of the results[7].
3.2.1. Effectiveness (RQ1): Our results demonstrate that the two most effective representations are those that consider the distance-based meta-features, producing the overall best results in all datasets considering both MicroF1 and MacroF1, with gains of up to 52%.

[8]For more results details (e.g. Figures and Confidence Interval of each fold) and impact of each step of the proposed pipeline, see Section 4 of our Master Thesis.

3.2.2. Efficiency Analysis (RQ2; RQ4): Comparing the efficiency of meta-features representations (without any additional step) and the traditional TF-IDF, the time of complete pipeline using MFs is higher than TF-IDF. However, in all cases, the other two proposed steps (SPA and SS) produced significant efficiency gains, individually or when applied in conjunction, compared to the MF versions. The SS step produces the largest time reductions compared to SPA, though the latter is also very efficient. The largest cost reductions are usually observed when combined, producing statistically significant effectiveness gains over TF-IDF with up to 9.7x processing time reductions.

3.2.3. Comparison with Embedding-Based Representations (RQ3): We run experiments with three embeddings-based representations: FisherVector, PTE and FastText. We standardize the use of the SVM as default classifier[9]. In terms of effectiveness, FastText is the worst among the embeddings representations, while PTE and FisherVector are more competitive. In any case, all embeddings-based representations are worse than (or at most tie with) TF-IDF with feature selection in all cases. Indeed, embeddings are never competitive when compared to MF representations. In all cases, there are significant losses. The embeddings results are even poorer in terms of efficiency: the overall times to generate them are 1.5x-31.1x slower than TF-IDF with feature selection. Compared to the overall times after applying our pipeline, the baselines' losses are even more significant – they reach up to 68.9x of slowdown.

3.2.4 Selective Sampling Applied to TF-IDF (RQ5): We evaluate the application of the SS step (Cover) on the TF-IDF representation after feature selection. We observe that the sampling time in this representation is *13.1x to 176.6x* slower than when applying it to MF representation. We also observe that SS applied to TF-IDF is *6.3x to 34.7x* slower than just using the original TF-IDF with feature selection. Simply put, as designed, the Cover is not suitable for application to the original TF-IDF representation before generating MFs. The MF denser representation is essential to guarantee the scalability of the SS step.

3.2.5 Impact on other Classifiers (RQ6): As representatives of state-of-the-art classifiers, we chose the Random Forest (RF) classifier and two recently proposed extensions of RF that have excelled in text classification tasks[10] : BROOF and BERT. We also chose two neural network architectures (NNs)[11]: Multilayer Perceptron (MLP); and Deep-MLP. First, we observe that the overall best results among all classifiers in all datasets, considering all metrics, are still produced by SVM. Regarding the RF and RF-based classifiers, we observe that using the three new steps of our proposal, we obtained better results in terms of effectiveness. The second important observation is that the application of the complete pipeline produced significant gains in terms of efficiency. Comparing with TF-IDF, we observe speedup gains between 1.3x-2.4x. Regarding the NNs, the application of both NN architectures to the original TF-IDF representation is too time-consuming. The complete pipeline with NN methods reduced the overall time considerably, with no effectiveness loss (in the case of MLP) and even gains over TF-IDF, confirming the enrichment of information brought by the MFs. To summarize, MF representations with sparsification and selective sampling make it possible to apply NN architectures to text classification tasks with potential effectiveness gains and much improved scalability.

[9]For more experimental setup details and cross-validation hyperparameters searching of embedding-based representations, see Section 4.2.3 and Table 4.6 of our Master Thesis.

[10]https://github.com/raphaelcampos/stacking-bagged-boosted-forests

[11]https://github.com/harvardnlp/sent-conv-torch

4. Conclusions and Future Work

We have provided evidence that perhaps even more important than the classifier algorithm is the adequate pre-processing of the data to achieve the best effectiveness results at the minimum cost. We introduced and orchestrated three new steps into the traditional text classification pipeline, which were shown to produce significant effectiveness gains, cost reductions (time), or both. By transforming the original textual representation, reducing dimensionality, increasing sparseness and "smartly" reducing the number of training instances, we achieved significant improvements in terms of effectiveness with decreased costs. We verified this by using a carefully designed, statistically rigorous experimental framework highlighting how each proposed pipeline step influences effectiveness. As future work, we envision the construction of AutoML solutions that could incorporate the proposed steps according to the datasets characteristics and goals of the task. We also plan to evaluate our proposal on additional datasets, classifiers, and configurations. We believe that our work has the potential to deeply change the way automatic text classification has been performed, with the potential for many new research opportunities.

References

B.-Naranjo, M., Martínez-Merino, L. I., and Rodríguez-Chía, A. M. (2021). A robust svm-based approach with feat. selection and outliers detection for classification problems. *Expert Systems with Applications*.

Canuto, S., Sousa, D. X., Gonçalves, M. A., and Rosa, T. C. (2018). A thorough evaluation of distance-based meta-features for automated text classification. *IEEE TKDE*.

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *IP&M*.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., and Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification. *IP&M*.

Cunha, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. (2018). A feature-oriented sentiment rating for mobile app reviews. In *WWW'18*.

Dacrema, M. F., Cremonesi, P., and Jannach, D. (2019). Are we really making much progress? In *RecSys*.

Kastrati, Z., Imran, A. S., and Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *IP&M*.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Neural Information Processing Systems NIPS'13*.

Mendes, L. F., Gonçalves, M., Cunha, W., Rocha, L., Couto-Rosa, T., and Martins, W. (2020). "Keep it simple, lazy" – MetaLazy: A new MetaStrategy for lazy text Classification. In *ACM CIKM'20*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *International Conf. on Language Resources and Evaluation LREC'18*.

Schoenfeld, B., Giraud-Carrier, C. G., Poggemann, M., Christensen, J., and Seppi, K. D. (2018). Preprocessor selection for machine learning pipelines. *CoRR*, abs/1810.09942.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *WSDM*.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Goncalves, M. (2020). CluHTM. In *ACL'20*.

Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S., Mourão, F., Salles, T., Rocha, L., and Gonçalves, M. A. (2018). Semantically-enhanced topic modeling. In *ACM CIKM'18*.

Zamani, H., Dehghani, M., Croft, W. B., Learned-Miller, E., and Kamps, J. (2018). From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *CIKM'18*.