Ensemble of Term Classification (ETC): Classifying Word Occurrences

Vítor Mangaravite[1]
Marcos André Gonçalves[1]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Departamento de Ciência da Computação (DCC),
Universidade Federal de Minas Gerais (UFMG),
Belo Horizonte, Brazil.

{vitormangaravite,mgoncalv}@dcc.ufmg.br

Resumo. *A classificação automática de texto no Processamento de Linguagem Natural (PLN) é a tarefa de prever classes para documentos textuais. Tradicionalmente, existem duas abordagens predominantes: modelos baseados em saco de palavras e modelos mais recentes baseados em sequência. Enquanto os modelos baseados em saco de palavras representam documentos considerando apenas a ocorrência de termos individuais, os modelos baseados em sequência levam em conta a ordem dos termos dentro do texto, até um comprimento máximo estabelecido. Embora os modelos de sequência baseados em aprendizado profundo tenham dominado o campo, abordagens baseadas em saco de palavras continuam a produzir resultados competitivos. No entanto, os métodos existentes geralmente exigem a construção de uma representação do documento e, em seguida, a previsão de sua classe, sem classificar especificamente as ocorrências individuais de termos dentro do conjunto de palavras. Essa lacuna na pesquisa serve como motivação para a tese proposta, que apresenta uma nova perspectiva para a classificação automática de texto. O objetivo principal é classificar cada ocorrência de termo dentro do saco de palavras e, em seguida, estimar a classe do documento, eliminando assim a necessidade de uma única representação oculta do documento. Como demonstrado pelos resultados obtidos, a abordagem proposta oferece maior interpretabilidade e eficiência na classificação de texto, ao abordar as limitações dos métodos existentes.*

1. Introduction

Automatic text classification (ATC) is critical in information systems, where topical categories is assigned to text units like documents, social media posts, and news articles. Sequence of Terms (SoT) and Bag of Words (BoW) are two common approaches. SoT models capture term semantics but may discard important terms since they consider only the first K tokens in the sequence. BoW models include all terms but mainly need more semantic understanding [Cunha et al. 2021].

Despite being state-of-the-art by most of the problems in NLP, SoT approaches face efficiency and interpretability challenges [Anelli et al. 2022]. Processing entire documents iteratively is used to slow performance, especially for lengthy texts. Recent SoT methods, based on Transformer architectures, have quadratic time complexity [Vaswani et al. 2017], further impeding classification speeds.

We propose the Ensemble of Term Classification (ETC) algorithm to address these issues. ETC takes a different approach by considering the classification of each unique term context within a document. It leverages term co-occurrence, assigning more weight to terms closely related to well-defined classes while ignoring uncertain terms.

ETC offers advantages in identifying complex term context combinations and providing a cost-effective, efficient, and scalable classification method. In this project, we aim to evaluate ETC against SoT and BoW algorithms, comparing their performance on various datasets using F1 micro and macro scores. Additionally, we will analyze prediction time cost to demonstrate ETC's efficiency and scalability. By addressing existing limitations, ETC has the potential to advance ATC and provide valuable insights into efficient and interpretable text categorization. The main contributions of this project include:

1. Introducing the Ensemble of Term Classification (ETC) algorithm as a novel approach to ATC.
2. Developing an efficient BoW algorithm suitable for classifying long documents.
3. Proposing a new representation for term context by utilizing the weighted co-occurrences of terms through a single (near-)Attention layer.

To guide our research and evaluate the effectiveness and potential advantages of ETC compared to baseline methods, we formulate two research questions based on the proposed approach:

- R1: To what extent does ETC demonstrate computational efficiency and maintain classification performance when processing large volumes of text data?
- R2: How effectively does ETC provide explanations for classification decisions by quantifying the importance of individual term contexts and their joint probabilities with labels?

Through rigorous experimentation and analysis, we will evaluate these hypotheses to gain a comprehensive understanding of the strengths and potential of the ETC approach. The results obtained will provide valuable insights into the performance, interpretability, scalability, and complementarity of ETC compared to the baseline methods, contributing to the ATC field.

1.1. Justification

The proposed project focuses on the ETC for automatic text classification, offering key justifications for its development. ETC provides inherent interpretability and explainability by classifying individual term contexts, enabling a better understanding of the classification process. Its explicit quantification of term context importance enhances transparency and allows users to comprehend the classification decisions, making ETC valuable in domains where interpretability is crucial.

Furthermore, ETC demonstrates scalability for large datasets, unlike traditional deep learning models that struggle with scalability [Cunha et al. 2021]. Operating efficiently on a term-by-term basis without complex mechanisms, ETC efficiently handles substantial volumes of text data without compromising performance. This scalability makes it a suitable solution for real-world applications dealing with big data scenarios.

In conclusion, the proposed project addresses important justifications, including interpretability, improved scalability, cost-efficiency, and bridging the gap in text classifi-

cation methods. These justifications underscore ETC's potential impact and significance in advancing ATC.

1.2. Related Work

In this section, we discuss the related work that has been conducted in the ATC field, particularly focusing on methods that are relevant to the proposed Ensemble of Term Classification (ETC) approach.

Supervised Latent Dirichlet Allocation (LDA) is a popular method for text classification that combines the probabilistic modeling of LDA with supervised learning. It aims to capture the latent topics within documents while incorporating labeled data for classification. In comparison to ETC, supervised LDA focuses on topic modeling and does not explicitly consider term contexts. ETC, on the other hand, operates at a granular level by classifying individual term contexts, providing interpretability and transparency in the classification process.

The BERT [Vaswani et al. 2017], RoBERTa [Liu et al. 2019], DistillBERT [Sanh et al. 2019], and XLNet [Yang et al. 2019] are models belonging to the family of Transformer-based architectures, which have significantly advanced the field of natural language processing (NLP). BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, DistillBERT, and XLNet are state-of-the-art models that leverage self-attention mechanisms to capture contextual information in texts. Unlike ETC, these models focus on learning comprehensive document representations to perform various NLP tasks, including text classification. ETC, however, takes a different approach by considering the classification of individual term contexts within documents. By doing so, ETC offers a more interpretable and granular classification process compared to the global representations generated by Transformer models.

In summary, the related work in ATC encompasses traditional approaches, like BoW and SoT. While traditional and deep learning methods have made significant contributions to the field, they have limitations in terms of interpretability and scalability. The proposed ETC method aims to address these limitations by considering the importance of term contexts and joint probabilities for text classification, offering new perspectives and opportunities for improving ATC techniques.

2. Formulation

In our proposed method, we consider the ATC dataset as $D = \{(d_i, l_i)\}$, which consists of N documents (d_i) and training labels (l_i) pairs. Each label l is one predictable topic in the L classes ($l \in L$) and each document d_i is the textual document. The problem can be generalized as predicting each label for all unseen documents D'.

In our formulation, the document is represented as a set of terms t_j and their respective frequencies $TF_{i,j}$ within the document (term frequency) and DF_j within the dataset (document frequency). We embed the term occurrences by encoding the term, term frequency (TF), and document frequency (DF) information into specialized one-hot-encode. The TF and DF encoders consider squared and logarithmic-scale encoding respectively, while the term encoder uses the traditional one-hot-encode. The embedding captures the joint influence of terms on the view of term frequencies and rarity representation.

The term context is a complex combination of the term's occurrences and their co-occurrences within the document. We approximate the probability as the ratio of its probability to the sum of probabilities of all term contexts within the document.

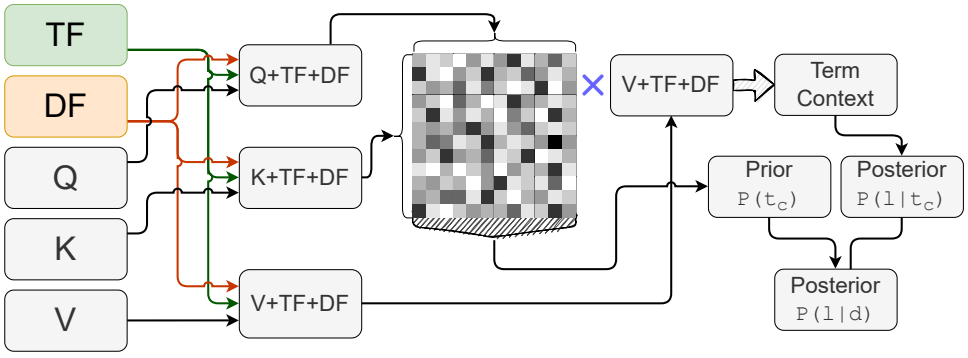

Figure 1. Diagram of ETC.

Our proposed framework is a probabilistic approach that represents the document/label posterior as the joint probabilities of term contexts and labels. It operates in a shared space of terms, their co-occurrences, and frequencies (TF and DF), which are also used for classification.

To quantify the probability space, we employ a (quasi-)Attention weighting model. This model approximates the probability of a term occurring in the context of another term based on their co-occurrence probabilities. The near-similarity between terms is calculated using a normalized distance measure and captures the contextual relevance of terms within the document. By considering the probabilities of term contexts, we can quantify the importance of each term context for classification.

Then, instead of directly estimating the document's class, our approach focuses on quantifying the importance of each term context for classification. We calculate the average joint probabilities of each term context with the label, reflecting the probability of the label given the term context and its prior.

These components together form the complete ETC framework, demonstrate in Figure 1. The framework leverages the joint probabilities of term contexts and labels to estimate the document class probabilities, providing an effective and interpretable approach to ATC.

3. Experiments

In this proposed project, we conducted experiments to evaluate the effectiveness of the Ensemble of Term Classification (ETC) approach for ATC. We used seven traditional classification datasets: 20 newsgroups (20ng), ACM, books, dblp, ohsumed, Web Knowledge base (webkb), and Web of Science (wos11967). Table 1 provides an overview of the main statistics for each dataset, highlighting their unique characteristics such as document sizes, number of classes, class balancing, and vocabulary size.

As baselines, we compared ETC against the best and most current sequence modeling methods widely considered state-of-the-art in ATC. These baselines included

dname	#Labels	#Docs	Class Ratio %	AVG doc len	AVG tokens per doc	#Vocab
20ng	20	18846	**94.3%**	243.63	**128.38**	**152843**
wos11967	**33**	11967	80.8%	193.64	113.59	57477
webkb	**7**	**8199**	31.6%	199.08	108.87	**26343**
ohsumed	23	18302	**27.6%**	185.03	100.6	51124
books	8	33594	85.1%	**260.88**	91.11	93356
dblp	10	**38128**	39.1%	139.41	84.4	67413
acm	11	24897	34.4%	**60.54**	**38.86**	56360

Table 1. The Table presents the following statistics: the column Class Ratio (%) provides the class ratio in percentage for each dataset. The class ratio indicates the imbalance of classes within the dataset ($AVG\_Len\_Classes/MAX\_Len\_Classes$); the column AVG doc len represents the average document length in each dataset; The column AVG tokens per doc indicates the average number of unique tokens per document in each dataset. The bold numbers represent the biggest values in the column and the bold-subscript values mean the smallest values in the column.

dname	20ng	acm	books	dblp	ohsumed	webkb	wos11967
ETC - Our approach	**0.904**$^{\bigcirc}$	0.678 $^{\bigcirc}$	0.87 $^{\bigcirc}$	0.805 $^{\bigcirc}$	0.713$^{\bigtriangledown}$	0.748$^{\bigtriangledown}$	**0.901**$^{\bigtriangleup}$
TFIDF+SVM [Cunha et al. 2021]	0.902 $^{\bigcirc}$	0.682 $^{\bigcirc}$	0.858$^{\bigtriangledown}$	0.801 $^{\bigcirc}$	0.704$^{\bigtriangledown}$	0.716$^{\bigtriangledown}$	0.837$^{\bigtriangledown}$
Albert [Lan et al. 2019]	0.596$^{\bigtriangledown}$	0.643$^{\bigtriangledown}$	0.843$^{\bigtriangledown}$	0.651 $^{\bigcirc}$	0.536$^{\bigtriangledown}$	0.695 $^{\bigcirc}$	0.677$^{\bigtriangledown}$
BERT [Vaswani et al. 2017]	0.875$^{\bigtriangledown}$	**0.704**$^{\bigcirc}$	**0.879**$^{\bigcirc}$	0.803 $^{\bigcirc}$	**0.767** $^{\bigcirc}$	0.82 $^{\bigcirc}$	0.835$^{\bigtriangledown}$
DistillBERT [Sanh et al. 2019]	0.864$^{\bigtriangledown}$	0.7 $^{\bigcirc}$	0.872 $^{\bigcirc}$	0.803 $^{\bigcirc}$	0.766$^{\bigcirc}$	0.816 $^{\bigcirc}$	0.834$^{\bigtriangledown}$
RoBERTa [Liu et al. 2019]	0.866$^{\bigtriangledown}$	0.696 $^{\bigcirc}$	0.87 $^{\bigcirc}$	0.808 $^{\bigcirc}$	0.695 $^{\bigcirc}$	**0.82**$^{\bigcirc}$	0.841$^{\bigtriangledown}$
XLNet [Yang et al. 2019]	0.87$^{\bigtriangledown}$	0.683 $^{\bigcirc}$	0.874 $^{\bigcirc}$	**0.81**$^{\bigcirc}$	0.763 $^{\bigcirc}$	0.742 $^{\bigcirc}$	0.847$^{\bigtriangledown}$

Table 2. The Table shows the micro F1 results obtained from the experiments. The best absolute values for each dataset are highlighted in bold. The symbols $^{\bigtriangledown}$ and $^{\bigtriangleup}$ indicate results that are statistically significantly lower or higher, respectively, than the best result (with a p-value < 0.05), while the symbol $^{\bigcirc}$ represents statistically equivalent results based on the t-student test.

BERT, RoBERTa, XLNet, DistillBert, Albert, and a traditional Bag of Words (BoW) approach using TFIDF with a linear SVM classifier. To evaluate the performance of each method, we employed the F1 micro and macro metrics, which summarize the accuracy of the models in terms of overall accuracy (micro) and accuracy per class (macro), accounting for dataset imbalances. To assess the statistical significance of the results, we applied ten-fold cross-validation strategies for each dataset. We used the t-student test with a confidence level of 95% to determine significant differences between the results.

3.1. Results

Tables 2 and 3 present the obtained F1 micro and macro for the results for ETC and the respective baselines, along with the statistical differences between the results. The experimental results revealed that ETC performed satisfactorily on two of the seven datasets. ETC emerged as the new state-of-the-art (SOTA) for datasets such as 20ng and WoS, which represent the two biggest datasets in the number of unique tokens (demonstrating the high accuracy on long datasets).

dname	20ng	acm	books	dblp	ohsumed	webkb	wos11967	Diff Best AVG^{STD}
ETC	**0.908**$^{\bigcirc}$	0.793 $^{\bigcirc}$	0.87$^{\bigtriangledown}$	0.824 $^{\bigcirc}$	0.77$^{\bigtriangledown}$	0.825$^{\bigtriangledown}$	**0.905**$^{\bigtriangleup}$	1.61%$^{2.29\%}$
TFIDF+SVM	0.904 $^{\bigcirc}$	**0.795**$^{\bigcirc}$	0.856$^{\bigtriangledown}$	0.819 $^{\bigcirc}$	0.764$^{\bigtriangledown}$	0.814$^{\bigtriangledown}$	0.842$^{\bigtriangledown}$	3.07%$^{2.82\%}$
Albert	0.61$^{\bigtriangledown}$	0.751$^{\bigtriangledown}$	0.841$^{\bigtriangledown}$	0.703 $^{\bigcirc}$	0.678$^{\bigtriangledown}$	0.811 $^{\bigcirc}$	0.684$^{\bigtriangledown}$	13.28%$^{9.69\%}$
BERT	0.879$^{\bigtriangledown}$	0.794 $^{\bigcirc}$	**0.879**$^{\bigcirc}$	0.821 $^{\bigcirc}$	0.816 $^{\bigcirc}$	0.876$^{\bigcirc}$	0.839$^{\bigtriangledown}$	**1.48%**$^{2.46\%}$
DistillBERT	0.868$^{\bigtriangledown}$	0.788 $^{\bigcirc}$	0.873 $^{\bigcirc}$	0.822 $^{\bigcirc}$	0.814 $^{\bigcirc}$	0.873 $^{\bigcirc}$	0.839$^{\bigtriangledown}$	1.88%$^{2.43\%}$
RoBERTa	0.871$^{\bigtriangledown}$	0.789 $^{\bigcirc}$	0.87 $^{\bigcirc}$	0.826 $^{\bigcirc}$	0.756 $^{\bigcirc}$	**0.876**$^{\bigcirc}$	0.845$^{\bigtriangledown}$	2.52%$^{2.74\%}$
XLNet	0.875$^{\bigtriangledown}$	0.783 $^{\bigcirc}$	0.874 $^{\bigcirc}$	**0.827**$^{\bigcirc}$	**0.818**$^{\bigcirc}$	0.832 $^{\bigcirc}$	0.851$^{\bigtriangledown}$	2.10%$^{2.21\%}$

Table 3. The Table shows the macro F1 results obtained from the experiments. The symbols and bold numbers mean the same as the Table 2.

dname	20ng	acm	books	dblp	ohsumed	webkb	wos11967
ETC	**777.87**$^{\triangle}$	**2815.9**$^{\triangle}$	**1101.4**$^{\triangle}$	**1577.3**$^{\triangle}$	**1145.0**$^{\triangle}$	**475.37**$^{\triangle}$	**1154.4**$^{\triangle}$
Albert	250.53$^{\triangledown}$	302.97$^{\triangledown}$	250.88$^{\triangledown}$	280.47$^{\triangledown}$	264.2$^{\triangledown}$	264.53$^{\triangledown}$	267.1$^{\triangledown}$
BERT	208.13$^{\triangledown}$	313.55$^{\triangledown}$	207.72$^{\triangledown}$	255.8$^{\triangledown}$	225.6$^{\triangledown}$	219.09$^{\triangledown}$	222.99$^{\triangledown}$
DistillBERT	283.47$^{\triangledown}$	524.03$^{\triangledown}$	283.08$^{\triangledown}$	379.25$^{\triangledown}$	315.98$^{\triangledown}$	303.78$^{\triangledown}$	315.96$^{\triangledown}$
TFIDF+SVM	145.7$^{\triangledown}$	390.4$^{\triangledown}$	119.88$^{\triangledown}$	89.385$^{\triangledown}$	151.3$^{\triangledown}$	255.92$^{\triangledown}$	196.84$^{\triangledown}$
RoBERTa	271.98$^{\triangledown}$	356.93$^{\triangledown}$	282.74$^{\triangledown}$	321.48$^{\triangledown}$	300.3$^{\triangledown}$	308.04$^{\triangledown}$	294.31$^{\triangledown}$
XLNet	200.09$^{\triangledown}$	230.32$^{\triangledown}$	201.59$^{\triangledown}$	217.21$^{\triangledown}$	207.99$^{\triangledown}$	207.25$^{\triangledown}$	210.75$^{\triangledown}$

Table 4. The Table shows the prediction velocity results obtained from the experiments. The best absolute values for each dataset are highlighted in bold. The symbols $\triangledown$ and $\triangle$ indicate results that are statistically significantly lower or higher, respectively, than the best result (with a p-value < 0.05).

In addition, ETC demonstrated statistical equivalence to the best-performing method in three of the seven datasets, and, despite the Books macro F1, the two datasets with worse performances represent the two most unbalanced datasets of the datasets studied, as shown in Table 1.

In summary, to show that ETC achieves statistically comparable results to the models investigated, we provide an inter-dataset post-analysis using the average macro F1 distances of each method for the best dataset outcome at Table 3 (column Diff Best). This metric allows us to determine how closely each method aligns with the best results across all datasets, with lower values indicating greater percentual proximity. Additionally, we present the standard deviation of each method's performance across all datasets to highlight the level of variability compared to the best result.

To evaluate the efficiency of ETC in comparison to the baselines, we conducted an analysis of prediction speed, as shown in Table 4. Prediction speed is measured as the number of documents that can be processed per second. Similar to the effectiveness analysis, the t-student test was utilized to assess statistical significance, comparing the prediction speeds per document across different folds.

The results indicate that ETC achieves state-of-the-art performance in terms of prediction speed. It outperforms the baselines, demonstrating faster processing capabilities. These findings highlight the efficiency and computational advantages offered by the ETC approach, making it a promising choice for applications requiring high-speed document classification.

4. Conclusion

Existing approaches, such as Sequence of Terms (SoT) and Bag of Words (BoW), have their limitations in terms of efficiency and interpretability. SoT models often suffer from time-based inefficiencies and may discard important terms, while BoW models lack semantic understanding. To tackle these challenges, we proposed the Ensemble of Term Classification (ETC) algorithm, offering a distinct approach to text classification. ETC utilizes term co-occurrence and incorporates a linear combination of classifications for each term context within a document. Our results demonstrate that ETC achieves statistically comparable accuracy while outperforming alternative methods with a speed improvement of up to 5.35 times without sacrificing performance.

In future research, it would be valuable to explore and demonstrate the interpretability properties of ETC in more depth. While ETC offers inherent interpretability by classifying individual term contexts, further investigations can be conducted to visualize

and explain the decision-making process of the algorithm. Additionally, investigating the integration of ETC with other techniques can lead to further improvements in performance and expand its potential applications in the field of automatic text classification.

References

Anelli, V. W., Biancofiore, G. M., De Bellis, A., Di Noia, T., and Di Sciascio, E. (2022). Interpretability of bert latent space through knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3806–3810.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., et al. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.