Fantastic Labels and Where to Find Them: Attention-Based Label Selection for Text-to-Text Classification

Michele Papucci[1,2,3], Alessio Miaschi[3] and Felice Dell'Orletta[1,3]

[1]Talia S.R.L., Pisa
[2]Università di Pisa
[3]ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Pisa

Abstract

Generative language models, particularly adopting text-to-text frameworks, have shown significant success in NLP tasks. While much research has focused on input representations via prompting techniques, less attention has been given to optimizing output representations. Previous studies found inconsistent effects of label representations on model performance in classification tasks using these models. In this work, we introduce a novel method for selecting well-performing label representations by leveraging the attention mechanisms of Transformer models. We used an Italian T5 model fine-tuned on a topic classification task, trained on posts extracted from online forums and categorized into 11 classes, to evaluate different label representation selection strategies. We've employed a context-mixing score called Value Zeroing to assess each token's impact to select possible representations from the training set. Our results include a detailed qualitative analysis to identify which label choices most significantly affect classification outcomes, suggesting that using our approach to select label representations can enhance performance.

Keywords

label selection, label representations, encoder-decoder, topic classification, attention mechanism

1. Introduction and Background

In recent years, generative language models have become increasingly prevalent for solving a wide range of NLP tasks. Among these models, the text-to-text paradigm has demonstrated significant success across numerous applications [2, 3, 4]. The text-to-text paradigm creates a unifying framework where each task is transformed to accommodate a textual input and output, resulting in a single abstraction capable of handling any task. Recently, the adoption and refinement of pre-trained Large Language Models (LLMs) have made this paradigm popular even in zero- or few-shot settings [5]. In these scenarios, most of the studies have focused on *prompting techniques* or *verbalizers*, i.e., how to better represent the input for the model, by specifying instructions or tasks. Few works have instead focused on how to better represent the output of the models. Among these, [6] designed different kinds of label representations and tested their impact on the T5 model on four classification tasks, showing that for most of these tasks, the performance was unaffected by the representations. Similarly, [7] showed that modifying the textual representation of the labels in a binary classification task (i.e. gender prediction) the performance of the IT5 model [8] does not change. On the contrary, shuffling the labels for a topic classification task leads to worse performance. By training several IT5 models with different label representations, [9] found that the textual representation of the label had a big impact on the model's discriminatory abilities for the same task of topic classification, especially for lower frequency classes. Nevertheless, an in-depth analysis focused on identifying correlations between model performance and several properties of the textual representations (e.g. the cosine distance between the encodings of the

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]
michele.papucci97@gmail.com (M. Papucci); alessio.miaschi@ilc.cnr.it (A. Miaschi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)
https://michelepapucci.github.io/ (M. Papucci); https://alemiaschi.github.io/ (A. Miaschi); http://www.italianlp.it/people/felice-dellorletta/ (F. Dell'Orletta)
0000-0003-4251-7254 (M. Papucci); 0000-0002-0736-5411 (A. Miaschi); 0000-0003-3454-9387 (F. Dell'Orletta)

representation and the original label name, the frequencies of the representations) yielded no significant insights on how to better chose these representations in order to maximize model performance.

Starting from these premises, in this work we propose a novel methodology for selecting label representations in a text-to-text classification scenario exploiting the potential of the attention mechanism of Transformer models. In fact, previous work showed that attention can be successfully employed in several scenarios, such as in the automatic identification of keyphrases from documents [10, 11], ontology alignment [12], document ranking [13] or semantic similarity [14]. Our purpose is to understand whether it is possible to define an automated approach for identifying a well-performing set of candidate labels in a classification task relying on a text-to-text model.

To investigate this, we conducted our experiments by fine-tuning the IT5 model on the topic classification task [15] using various label representations. Specifically, we tested different approaches for selecting candidate labels relying on *Value Zeroing* [16], a context-mixing score based on the attention mechanism aimed at quantifying the contribution each context token has in determining the final representation of a target token. Moreover, we performed a thorough qualitative analysis to determine which labels have the most substantial impact on the improvement or decline of classification results.

Contributions In this paper we: i) present a novel technique for label representation selection based on the attention mechanics of Transformers models. We tested three different configurations and found that one shows promising results in finding the best possible representations to maximize performances; ii) show an in-depth qualitative analysis of the chosen representation, with the intent to find usable correlations to improve the performance of our label representation selection technique.

2. Our Approach

When employing a text-to-text model for classification tasks, the class names must be represented as specific sequences of tokens (hereafter **label representations**) that the model outputs to assign an input to a particular class. We aim to find a set of suitable label representations that maximize the model's performance.

To do so, we hypothesize that we can use the attention mechanism of the model to find suitable representations for each class inside the training set of the target task. Particularly, we look at which tokens were the most salient for building the vectorial representations of *important tokens* in the post using Value Zeroing. We tested three different ways to select the *important tokens* inside the posts. First, we tried looking at the tokens that were used to build the representation of the End-of-Sentence special character of T5 </s> (EOS). Then we also tried to append class-related tokens to the end of the posts. The idea was to inject class-related words into the posts to see which original tokens from the posts were useful in building them:

- In the **Appended Label** method, we define p as the translation of the original class names[1], e.g. the post *"Che giornata indimenticabile... è passato proprio tanto tempo!</s>"* from category SPORTS, becomes: *"Che giornata indimenticabile... è passato proprio tanto tempo! Sport</s>"*;
- In the **Appended Label with Prompt** method, we provide the model additional context, by defining p as: *La frase precedente appartiene alla categoria* (English translation: *The previous post belongs to the category of*) followed by the original class name translated, e.g. the post *"Che giornata indimenticabile... è passato proprio tanto tempo!</s>"* from category SPORTS, becomes: *"Che giornata indimenticabile... è passato proprio tanto tempo! La frase precedente appartiene alla categoria Sport</s>"*

Formally, let $S \in D$ as one of the training posts in the dataset D. Each post S is tagged with one of the classes $c \in C$ where C is the set of the possible topics. The posts are tokenized using the provided IT5-trained tokenizer T. For each post S, we injected a series of tokens p tokenized with T. The

[1]List of translated labels: *anime*, *automobilismo*, *bicicletta*, *sport*, *natura*, *metal detector*, *medicina*, *celebrità*, *fumo*, *intrattenimento* and *tecnologia*.

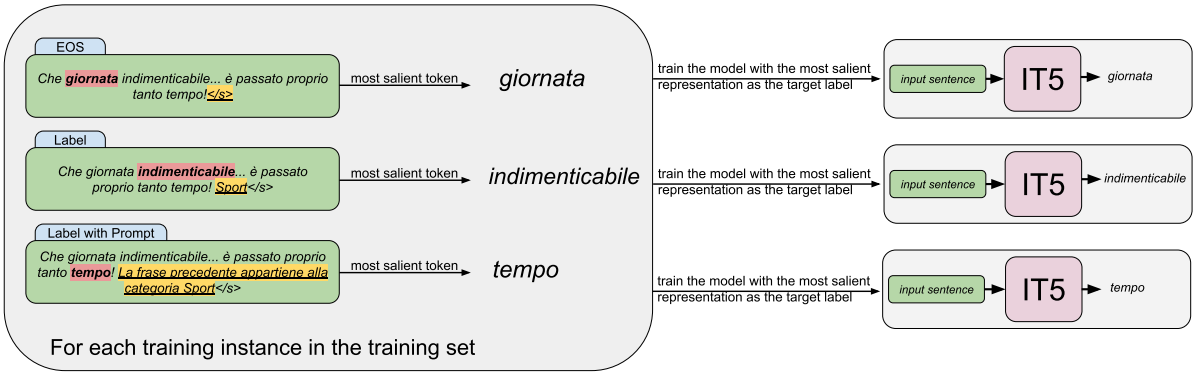

Figure 1: Overview of the proposed approach. First, we extract the most salient token from each training instance using Value Zeroing. Then, according to their scores, we use the best ones as label representations.

objective is to study which tokens from the original post S are more salient for the model to build the representation of the tokens in p.

As explained before, the difference between the three methods is how p is defined: in the EOS method $p = </s>$, in the Appended Label method p is equal to the appended and translated class name, and in the Appended Label with Prompt method p is equal to the predefined prompt completed with the translated class name.

After injecting p in each $S \in D$ we pass each post in inference through our modified implementation of IT5, whose Encoder is able to calculate the Value-Zeroing matrix (see Section 3.1). Then, we define as a candidate label representation l the token $s \in S$ that obtained the highest Value-Zeroing score with respect to the tokens in p:

$$l = argmax(value\_zeroing\_score(S, p))$$

By doing so we obtain, for each post, the most important token whose embedding vector is used to construct the representation of p[2]. After doing it for the whole dataset, we obtain, for each category $c \in C$, a list of representations R_c. R_c contains n tuples r_i equal to the number of posts in the dataset D tagged with c. Each tuple is composed by the candidate label representation l and the Value-Zeroing score vz it obtained with respect to p, $R_c = [(l_1, vz_1), ..., (l_n, vz_n)]$.

Since some of these representations may be duplicates (i.e. the same representation l has been chosen from multiple posts), we decided to aggregate those representations, in a way that rewards their higher frequency count. We aggregate all the tuples that have the same representation l and sum together their Value-Zeroing score vz creating a single element in the R_c list. After doing these aggregation steps for all categories $\forall R_c : c \in C$, we have, for each category c, a set of representations R_c that we sort based on the vz value of the tuples in descending order, obtaining a ranked Representation Set.

Finally, we define a set of representations E_i, called the Representation Set of rank i where, for each category c, we have the i ranked representation r_i in R_c. E.g. in the set E_0, for each category, we have the best-ranked representation, while in the set E_{10}, for each category we have the representation that ranked 11th.

An overview of our approach is illustrated in Figure 1.

[2]Since Transformers' tokenizers split tokens in multiple subtokens, to obtain the full word, we reconstruct it by reconnecting all the tokens that are part of the word that the token with the highest Value-Zeroing score is from. The Value-Zeroing score we consider for the full word is the one of the token that was selected. We decided to avoid aggregating the score of the full word in any way, because that could reward or punish multi-token words .

3. Experimental Setting

We tested our approach to solving a topic classification task by training our models on forum posts categorized into 11 classes. We tested all three previously presented selection methods and evaluated their performances: we used as the target output the first ten best-performing Representation Sets $E_0, E_1, ..., E_9$ for all three methods, training ten models for each, for a total of 30 trained models. Then, having assessed that the most promising strategy was the EOS method, we trained 100 models using the Representation Sets $E_0, ..., E_{99}$ extracted with the EOS method to study the effectiveness of this+ approach.

In the following sections, we detail how the Value Zeroing technique works (Sec. 3.1), we present the data, the model and the evaluation methods used in our experiments (Sec. 3.2 and 3.3).

3.1. Value Zeroing

Value Zeroing [16] draws inspiration from traditional interpretability techniques, where the influence of a feature (in this case, a token representation) on the model's output is extracted by removing that feature from the input, i.e. feature importance methods [17]. Since deleting a word from a sentence, without changing the semantics of it, is either challenging or impossible, the method opts to *eliminate it* during the Attention computation of the considered layer, by *zeroing* its value vector, i.e. setting each element in the vector to 0. Inside the Self-Attention layer of a Transformer, for each Attention head h, the input vector $\mathbf{x}_i$, for the i^{th} token in the sequence is transformed in three distinct vectors through the use of different sets of weight: the Query vector $\mathbf{q}_i^h$, the key vector $\mathbf{k}_i^h$ and the Value vector $\mathbf{v}_i^h$. The context vector $\mathbf{z}_i^h$ for the i^{th} token of each Attention head is generated as a weighted sum over the Value vector:

$$\mathbf{z}_i^h = \sum_{j=1}^{n} \boldsymbol{\alpha}_{ij}^h \mathbf{v}_j^h \tag{1}$$

where $\boldsymbol{\alpha}_{ij}^h$ is the raw Attention weight assigned to the j^{th} token and computed as a Softmax-normalized dot product between the corresponding Query and Key vectors. In Value-zeroing Equation 1 is changed by replacing the Value vector associated to j with a zero vector $\mathbf{v}_j^h \leftarrow \mathbf{0}, \forall h \in H$, where the context vector for the i^{th} token is being computed. This provides a new representation $\mathbf{x}_i^{\neg j}$ that has excluded j. By comparing the original representation $\mathbf{x}_i$ with this new one, usually by means of a pairwise distance metric, we obtain a measure of how much the output representation is affected by the exclusion of j. In our experiment, we chose the *cosine distance* as a distance metric:

$$\mathbf{C}_{ij} = cs(\mathbf{x}_i^{\neg j}, \mathbf{x}_i) \tag{2}$$

Computing Equation 2 for each token i, j generates a **Value-Zeroing Matrix $\mathbf{C}$** where the value of cell $\mathbf{C}_{ij}$ in the map indicates the degree to which the i^{th} token is dependent on the j^{th} to form its contextualized vectorial representation.

For our experiments, we modified the implementation of T5 in the Python transformers library[3] such that the model's encoder can calculate the Value-Zeroing Matrix $\mathbf{C}$. In particular, we look at the section of the matrix $\mathbf{C}_{n_s:n_p,0:n_s}$ where n_s is the number of tokens in the original sentence and n_p is the number of tokens that compose the appended specially placed tokens (See 2 for how we chose these tokens). This section of $\mathbf{C}$ illustrates how each original token in the sentence contributes to the vectorial representation of the appended tokens.

3.2. Data

We relied on posts extracted from TAG-IT [15], the profiling shared task presented at EVALITA 2020 [18]. The dataset, based on the corpus defined in [19], consists of more than 18,000 posts written in

[3]The modified class is *T5ForConditionalGeneration* available in https://github.com/huggingface/transformers/blob/main/src/transformers/models/t5/modeling_t5.py. To do so, we adapted the original Value Zeroing implementation for the BERT transformer modelling class: https://github.com/hmohebbi/ValueZeroing.

Categories	# Data	# Training	# Test
Anime	3,972	2,894	1,078
Auto-Moto	3,783	2,798	985
Bikes	520	365	155
Celebrities	1,115	754	361
Entertainment	469	354	115
Medicine-Aesthetics	447	310	137
Metal-Detecting	1,382	1,034	348
Nature	516	394	122
Smoke	1,478	1,101	377
Sports	4,790	3,498	1,292
Technology	136	51	85
All	18,608	13,553	5,055

Table 1
Dataset statistics.

Italian and collected from different blogs. Each post is labeled with three different labels: age (binned into 5 classes) and gender (male or female) of the writer, and topic (11 classes).
Since previous works have shown that tasks that are solved through the use of lexical and semantic information benefit the most from a well-chosen label representation [7, 9], we have decided to focus only on the Topic classification task. Moreover, to have comparable results with previous studies, we used the same dataset configuration used in [9]. This setting is different from how the original task was defined in [15]: instead of predicting the label of a given collection of texts (multiple posts), we fine-tuned our model to predict the topic from each single post and, since a fair amount of posts was quite short, we removed the posts shorter than 10 tokens. At the end of this process, we obtained a dataset consisting of 13,553 posts as the training set and 5,055 posts as the test set. The distribution of posts according to each label is reported in Table 1.

3.3. Model and Evaluation

We used the T5 base version pre-trained on the Italian language, i.e. IT5[4]. In particular, the model was trained on the Italian sentences extracted from a cleaned version of the mC4 corpus [20], a multilingual version of the C4 corpus including 107 languages.

Models' performances on the topic classification task were computed using the F-Score on the test set. To evaluate the capability of our selection method to find suitable labels, we trained up to 100 models with 100 different Representation Sets. Each of these sets was composed of representations chosen by our method and was ranked based on its prediction, from the set predicted as the best (Rank 0) to the set predicted to be the worst (Rank 99). We then calculated the Spearman correlation between the set's ranking, and the obtained F-Score using that set. If our method can reliably predict the best representation to maximize performance, we expect a correlation between the ranking and the model performances. We used the traditional approach of using translated class names for classification as our baseline.

4. Results

As a first step, we evaluated the first ten Representation Sets $E_0, ..., E_9$ from each of the three tested methods to assess their potential in predicting the most effective representations. Figure 2 reports the scatter-plots showing the F-scores obtained on the test set by each model according to the 10 Representations Sets. As we can notice, the first two methods, Appended Label and Appended Label with Prompt, don't show any particularly interesting trends. The first one has a slightly negative coefficient and a Spearman Correlation of 0.03 with a p-value of 0.934. With such a low correlation

[4]https://huggingface.co/gsarti/it5-base.

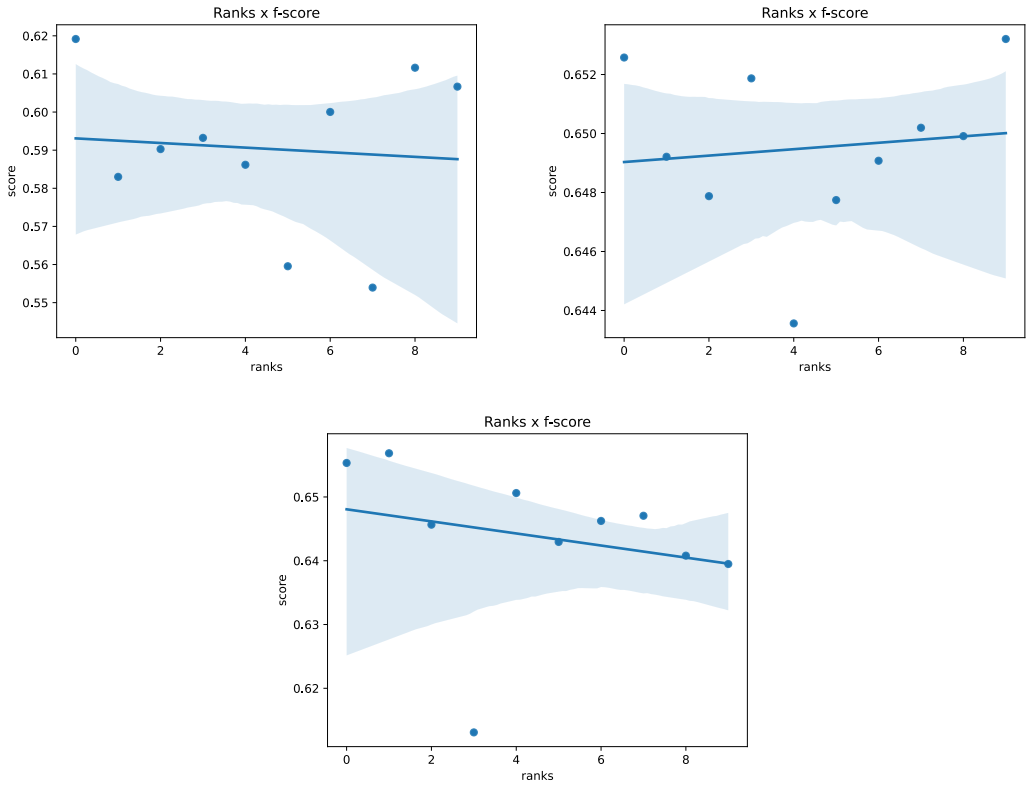

Figure 2: Scatter-plots with regression lines where each point is a model. On the y-axis we have the weighted F-Score on the test set and on the x-axis we have the rank of the Representation Set used to train it. On the top-left we have the Label Appended method, then on the top-right the Appended Label with Prompt method and on the bottom the EOS method.

value and high p-value we can't reject the null hypothesis and the obtained trend is probably random. The same can be said for the second method too, where we have a slightly positive trend, with a Spearman correlation of 0.151 with a p-value of 0.67. On the contrary, the third method shows a more pronounced negative trend ($Spearman = -0.552$), i.e. as the rank increases the performance of the models tends to decrease. Although the p-value of the correlation is below the standard cut-off threshold ($p - value = 0.098$), we decided to use the EOS method for testing with a total of 100 representation sets. Before proceeding, we removed from the original dataset the posts belonging to TECHNOLOGY. This was done since for this class we extracted only 23 sets, due to the small number of samples in the training set. After removing this class, we evaluated the method with the rest of the categories training 100 models with the first 100 ranked Representations Sets.

Correlation results are reported in Figure 3, while Table 2 shows the performances of the models obtained with the representation set of rank 0, the best performing model (Ranked 20th), and the worst performing model (Ranked 95th) along with the baseline (A IT5 trained with the original class label translated into Italian). As we can see, the negative trend between models' performance and Representation Sets observed previously can still be noted, although less pronounced ($Spearman = -0.314$, $p - value = 0.001$). In terms of classification scores, we obtained a difference of 0.05 in terms of F-score between the best-performing model obtained by rank 20 (0.68), and the worst-performing one obtained by rank 95 (0.63). Although general conclusions about the method cannot be drawn, it appears that, in this setting, selecting labels from the training set using Attention attribution techniques, such as Value-Zeroing, effectively identifies keywords with meaningful semantic connections that IT5 can leverage to achieve higher performance.

Interestingly, the lowest-performing model using the EOS method achieved the same F-score (0.63) as the baseline method, i.e. the standard approach of using translated class names. From this perspective,

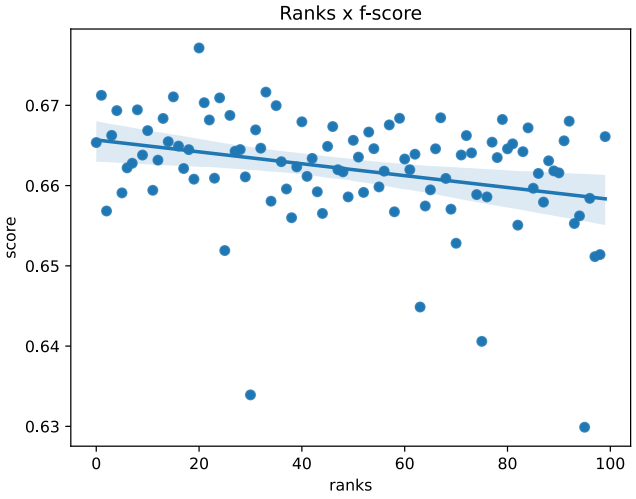

Figure 3: Scatter plot for the first 100 Representation Set extracted with the EOS method from the training set where the low-frequency TECHNOLOGY class was removed.

Representation Set	F-Score
0 (First Set)	0.66
20 (Best Performing Set)	0.68
95 (Worst Performing Set)	0.63
Baseline (Trained with original class names)	0.63

Table 2
F-scores for different representation sets.

the EOS method demonstrates superiority over the standard approach: in fact, the model trained on the Representation Set ranked 0, identified by the EOS method as the best set, achieved an F-score of 0.656. While this is not the highest score produced by the EOS method, it still outperforms the standard approach. A possible explanation of the effectiveness of the EOS method could be that for building the </s> character, the Encoder of the model uses particularly informative words that we can leverage if used as label representation. The role of the EOS character and other similar characters that are used for modeling purposes, like the *[CLS]* character in BERT-like models, is to be used as input for the final Language Modeling classifier. That pushes the model during the pre-training phase to learn to construct a representation of such a token that summarizes all the relevant information in the sentence that is needed to complete the language modeling task [21]. So, by taking the highest Value-Zeroing score for constructing </s>, we find tokens that are usually very contextually informative to the language modelling task and contain a lot of useful information. It's likely, then, that this information is also useful during the fine-tuning phase, to construct that lexical connection between input sentences and output classes. Moreover, when using the other two techniques, we focus on injected tokens that are often appended without sufficient context to justify their presence at the end of the post. It may be that appending tokens to the end of the post may change the semantics of the sequence too much. The first method, which simply appends the label to the end of the post, often creates scenarios where the word appears to be out of place. The same applies to the second method, but thanks to the prompt, this effect is less noticeable. This effect may also be the reason why the first method is the worst performing one, while the Appended Label with Prompt method achieves F-Scores almost as high as the EOS one but without showing any useful correlation between the chosen representation and the model F-Score, thus not being usable as a Label Representation Selection method.

Figure 4 shows the variation in F-Scores obtained for each class. As we can observe, there is a quite low degree of variance between the classes, in contrast with the results obtained in [9], which used the same dataset and task, but represented the classes using 10 human-selected representations and 90

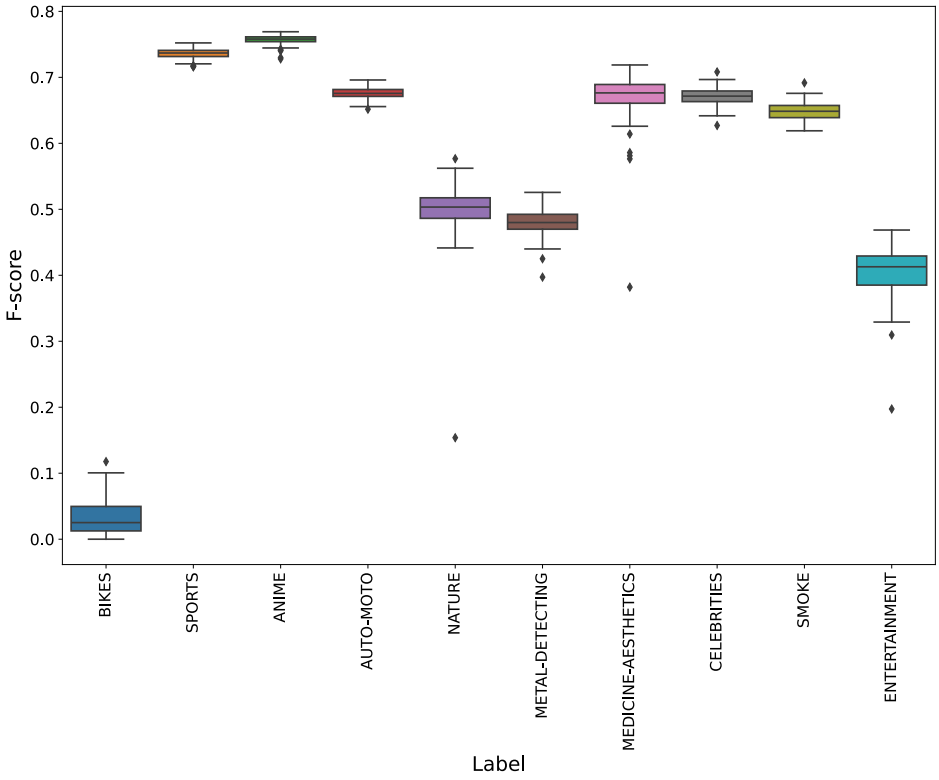

Figure 4: Boxplot for each Category to show the variations in F-Score obtained using the various Representation Sets.

Class	Best Performing Representation Set (Rank 20) F-Score: 0.68	Worst Performing Representation Set Rank (95) F-Score: 0.63
BIKES	risolvo	temperatura
SPORTS	schedina	decidesse
ANIME	troverai	principiante
AUTO-MOTO	premuto	abbassato
NATURE	gippi	causarne
METAL-DETECTING	cosa	pistolina
MEDICINE-AESTHETICS	capelluto	soffermarmi
CELEBRITIES	ilaria	scherzo
SMOKE	eccola	piaciuto,proviamo
ENTERTAINMENT	origini	dragonette

Table 3
Table showing the representations for the best performing and worst performing set in the first 100 Representation Sets extracted from the training set where the posts of the TECHNOLOGY class were removed.

randomly selected ones. This is especially pronounced for the lower frequency categories, where high F-score scores also correspond to lower variance. For instance, in contrast to the results reported by [9], where the MEDICINE-AESTHETICS class exhibited numerous outliers with F-scores dropping to as low as 0, our selection method does not encounter such extreme variations. Even when accounting for outliers, the performance of the class remains relatively stable, with F-scores that are acceptable even in the worst-case scenario. A similar trend is observed for the ENTERTAINMENT class.

Class	F-score x Frequency	Rank x Frequency	F-Score x TF-IDF	Rank x TF-IDF	F-Score x Subtoken Length	Rank x Subtoken Length
BIKES	-0.080	0.080	-0.118	0.007	-0.031	0.029
SPORTS	0.138	-0.307*	0.213*	-0.499*	-0.166	0.303*
ANIME	0.125	-0.101	0.283*	-0.408*	0.059	-0.071
AUTO-MOTO	0.147	-0.346*	0.081	-0.290*	-0.408*	0.126
NATURE	0.049	-0.026	0.148	-0.159	0.113	0.132
METAL-DETECTING	-0.053	-0.128	0.146	-0.204*	-0.130	0.181
MEDICINE-AESTHETICS	0.152	-0.149	0.025	-0.067	-0.116	0.074
CELEBRITIES	-0.182	-0.014	0.031	-0.329*	-0.071	0.210*
SMOKE	0.030	-0.034	0.080	-0.388*	-0.090	0.127
ENTERTAINMENT	-0.099	-0.005	-0.103	0.004	0.116	0.048

Table 4
Correlations between frequencies, TF-IDFs, number of subtokens and F-scores and the Representation Sets rank.

4.1. Qualitative Representations Analysis

To have a deeper understanding of the effectiveness of the approach, we performed a more qualitative analysis to determine which labels have the most substantial impact on the improvement or decline of the classification accuracy. Table 3 reports the representations for each class obtained with the best and worst performing models. As we can see, and in line with previous work [6, 9], it would seem there are no clear patterns that could justify why certain words work better than others. Focusing on the best-performing set, the only two words that are somehow related to their class seem to be *schedina* for SPORTS, referring to the betting ticket used to bet for sports games, and the proper noun *ilaria* for CELEBRITIES. For the worst representation, the only representation that can fit in its corresponding domain (ENTERTAINMENT) is again a proper noun: *dragonette*, which is the name of a Canadian band. Another interesting case is *piaciuto,provato*, which was treated as a single word by the IT5 tokenizer giving the missing space after the comma. While our aggregation method for multiply selected tokens also rewarded the frequency, from the best-performing representations, only four had been chosen as the most salient token in the text multiple times: *schedina* (3 times), *troverai* (2 times), *premuto* (2 times), *gippi* (2 times). This could mean that we should re-evaluate how important frequency is, and maybe change the aggregation method to something that doesn't reward the frequency as much.

To better understand the role of the representations frequencies in the training set we computed both the raw frequency of each representation in the whole dataset and the TF-IDF of the representations. We then calculated the Spearman Rank between the frequencies, the TF-IDFs, and the number of subtokens of the representations against the obtained F-Score and the Representation Sets rank the representations are in. The TF-IDF has been calculated by considering all the documents of a single category as a single document, and the documents' length has been calculated as the total number of tokens (document lengths are reported in Appendix A). As we can observe (Table 4) representation frequency does not correlate to any class with the obtained F-Score. This, again, confirms that the role of the absolute frequency of a certain term in the training set doesn't seem to have any positive or negative effect on the ability of the model to use such representation for its classes. However, by using the aggregation method that rewarded frequency mentioned in Sec. 2, we can see that for some classes the more frequent a word is, the more likely it is to be placed at a better rank. (Rank x Frequency column in Table 4). In particular, for two classes (SPORTS and AUTO-MOTO) more frequent representations had a higher chance to be placed in the best ranks. This could mean that the most informative words are frequently the same in these particular categories. Focusing on the TF-IDFs correlations, we can notice two negative statistically significant correlations with the F-score: SPORTS and ANIME. This is probably due to the fact that in-domain words, that don't appear as often in the other categories, had a slightly positive impact on performances. Moreover, we noticed that these two categories are also those for which our model utilized several domain-specific words. In fact, the first ten ranked representations for the two categories are mostly domain-specific:

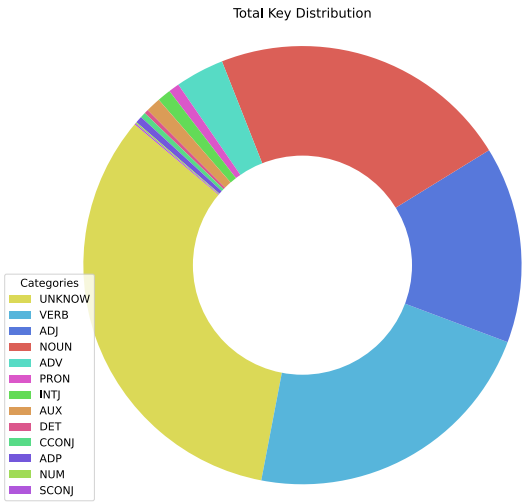

Figure 5: Distribution of the Parts-of-Speech across all the representations extracted using the EOS method.

- for SPORTS: *campionato*, *gol*, *pareggio*, *centrocampo*, *milan*, *juventus*, *atalanta*, *tifosi*, *trequartista* and *derby*;
- for ANIME: *streaming/download*, *graffio*, *manga*, *pokémon*, *pokemon*, *ko*, *morso*, *pokèmon*, *cmq*, *drago*[5].

Again, the correlation between the TF-IDF and the Representations' rank was to be expected, based on the method we've used to aggregate the representations. We've also empirically noticed that the method we've used to extract the representations was keen on choosing domain-specific, low-frequency words. This is why often it chooses typos and similar words with errors in them. This is probably because low-frequency words are usually *full words* with much more semantic value in them, and by being domain-specific they carry high contextual information, useful for constructing the other tokens' representations. This could explain why the TF-IDF, which is a metric that is specifically built to find such words, correlates so highly and significantly with the extracted words' ranks.

Since Transformer models tokenize text by splitting them into subwords, we also tried to understand whether there is any correlation between both the F-score and the Representation rank with the number of subwords of the representations. From our results, we can see that subword length doesn't seem to affect the model's performance, nor does our selection technique seem to prefer words that are split into more or less subwords. The only two exceptions are AUTO-MOTO, where a higher number of subwords leads to a decrease in performances, and SPORTS, where our model seemed to place words with a higher subword number in lower places in the ranking system.

Finally, we investigated the impact of the Part-of-Speech (PoS) associated with the representations, both globally (See Figure 5) and on a per-class basis (Class-based distribution are reported in Appendix B). The PoS are extracted from an Italian Word Form Vocabulary developed by the Institute for Computational Linguistics (ILC) of the National Research Council of Italy (CNR), which contains all the word forms and their possible POs from the Italian language. As we can see from Figure 5, the most frequent PoS are UNKNOWN, VERB, NOUN, and ADJ. The class UNKNOWN contains the words that are not found in the Word Form Vocabulary, and these usually consist of typos, English words, proper nouns, etc. and are going to be seen more in detail for each category. The categories with the highest number of UNKNOWNs are ENTERTAINMENT, CELEBERITIES, ANIME, and SPORTS:

[5] *morso*, *graffio* and *ko* are all domain-specific words in the settings of the popular anime, cartoon and video-game Pokémon, with the first two being moves and the latter being a specific status.

- in ENTERTAINMENT, the majority of the UNKNOWNs are typos (e.g. *cioe* instead of *cioè*), abbreviations (e.g. *nnt* instead of *niente*), words with an increased vocal length in the last character (e.g. *iniziaaaaaa* instead of *inizia*) or english words (e.g. *wish*);
- in CELEBRITIES, the majority are proper nouns (e.g. *alessia*, *mirco*, *federica*, etc.) and typos;
- in ANIME, the majority are proper nouns of video games or tv shows characters (e.g. *pokémon* or *charmender*) or Japanese words (e.g. *manga*);
- in SPORTS, the majority are proper nouns of soccer teams or players (e.g. *milan*, *juventus*, *higuain*, etc.) or match names composed by multiple teams or nation names (e.g. *italia-uruguay* or *brasile-olanda*) that our system didn't split since they didn't contain any spaces.

We also noted that for BIKES, NATURE, and AUTO-MOTO more VERBs are chosen instead of NOUNs, while for METAL-DETECTING, SPORTS, and SMOKE is the contrary. That being said, all the Parts-of-Speech seem reasonably distributed and it seems that no particular one is preferred by the method when choosing representations from the training set.

5. Conclusion

In this work, we presented a novel technique for reliably choosing label representation in text-to-text classification scenarios. This novel technique, based upon an Attention attribution technique called *Value Zeroing*, provides a set of labels used to represent the class names for a text-to-text model. We tested the approach on a Topic Classification task using IT5, an Italian pre-trained T5 model, by training 100 different models with 100 sets of representation chosen this way. We found that choosing representation with Value Zeroing and ranking them based on its value, leads to a useful correlation with the trained model's scores. Moreover, we noticed that choosing representation this way, leads to better average performances and lower variance in performance, against both human- and random-chosen representations [9]. Compared to the standard approach of using the class names directly as their representation (in this case, by also translating them to Italian) our method performed better, and even the worst-performing Representation set matched the standard approach.

We also conducted an in-depth analysis to understand whether either the performance of the model or our rankings were related to some simple statistics (frequency, TF-IDF, and the number of subtokens of the representations). Results showed some statistically significant correlations, especially when focusing on the TF-IDFs of the representations. We also found no interesting trend among the Parts-of-Speech of the representation chosen this way. While NOUNs and VERBs were the most popular, there weren't any interesting findings, and some distributions suggest that the chosen representations are usually low-frequency in-domain words for that class.

In conclusion, our findings highlight again that the choice of label representations isn't trivial and has an important impact on text-to-text classification performances and our technique seems to be a way to find a good solution for the label representation selection task.

Future research should focus on applying this technique to different kinds of tasks, primarily on those tasks where lexical and semantic clues from the text are essential in solving the task. Also, other aggregation methods should be tested, reducing the impact of the selection frequency, which showed not to be an important factor in the fine-tuned models' performances.

Acknowledgments

This work has been supported by:

FAIR - Future AI Research (PE00000013) project under the NRRP MUR program funded by the NextGenerationEU.

TEAMING-UP - Teaming up with Social Artificial Agents project under the PRIN grant no. 20177FX2A7 funded by the Italian Ministry of University and Research.

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] V. Aribandi, Y. Tay, T. Schuster, J. Rao, H. S. Zheng, S. V. Mehta, H. Zhuang, V. Q. Tran, D. Bahri, J. Ni, et al., Ext5: Towards extreme multi-task scaling for transfer learning, in: International Conference on Learning Representations, 2021.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[4] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. Le Scao, A. Raja, et al., Multitask prompted training enables zero-shot task generalization, in: The Tenth International Conference on Learning Representations, 2022.

[5] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[6] X. Chen, J. Xu, A. Wang, Label representations in modeling classification as text generation, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, 2020, pp. 160–164.

[7] M. Papucci, C. De Nigris, A. Miaschi, F. Dell'Orletta, Evaluating text-to-text framework for topic and style classification of italian texts, in: Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022), 2022.

[8] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.

[9] M. Papucci, A. Miaschi, F. Dell'Orletta, Lost in labels: An ongoing quest to optimize text-to-text label selection for classification, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), Proceedings of the 9th Italian Conference on Computational Linguistics, Venice, Italy, November 30 - December 2, 2023, volume 3596 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3596/paper39.pdf.

[10] H. Ding, X. Luo, AttentionRank: Unsupervised keyphrase extraction using self and cross attentions, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1919–1928. URL: https://aclanthology.org/2021.emnlp-main.146. doi:10.18653/v1/2021.emnlp-main.146.

[11] B. Kang, Y. Shin, SAMRank: Unsupervised keyphrase extraction using self-attention map in BERT and GPT-2, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 10188–10201. URL: https://aclanthology.org/2023.emnlp-main.630. doi:10.18653/v1/2023.emnlp-main.630.

[12] V. Iyer, A. Agarwal, H. Kumar, VeeAlign: Multifaceted context representation using dual attention for ontology alignment, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10780–10792. URL: https://aclanthology.org/2021.emnlp-main.842. doi:10.18653/v1/2021.emnlp-main.842.

[13] Z. Li, C. Tao, J. Feng, T. Shen, D. Zhao, X. Geng, D. Jiang, FAA: Fine-grained attention alignment for cascade document ranking, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1688–1700. URL: https://aclanthology.org/2023.acl-long.94. doi:10.18653/v1/2023.acl-long.94.

[14] H. Yamagiwa, S. Yokoi, H. Shimodaira, Improving word mover's distance by leveraging self-attention matrix, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 11160–11183. URL: https://aclanthology.org/2023.findings-emnlp.746. doi:10.18653/v1/2023.findings-emnlp.746.

[15] Cimino, Dell'Orletta, Nissim, Tag-it – topic, age and gender prediction, EVALITA (2020).

[16] H. Mohebbi, W. Zuidema, G. Chrupała, A. Alishahi, Quantifying context mixing in transformers, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3378–3400. URL: https://aclanthology.org/2023.eacl-main.245. doi:10.18653/v1/2023.eacl-main.245.

[17] S. Mishra, S. Dutta, J. Long, D. Magazzeni, A survey on the robustness of feature importance and counterfactual explanations, 2023. URL: https://arxiv.org/abs/2111.00358. arXiv:2111.00358.

[18] V. Basile, M. Di Maro, D. Croce, L. Passaro, Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, volume 2765, CEUR-ws, 2020.

[19] A. Maslennikova, P. Labruna, A. Cimino, F. Dell'Orletta, Quanti anni hai? age identification for italian., in: CLiC-it, 2019.

[20] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[21] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of BERT's attention, in: T. Linzen, G. Chrupała, Y. Belinkov, D. Hupkes (Eds.), Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Florence, Italy, 2019, pp. 276–286.

A. Documents Size for TF-IDFs

Class	Document Size in Tokens
ANIME	154.347
BIKES	14.078
SPORTS	163.142
AUTO-MOTO	118.859
NATURE	21.595
METAL-DETECTING	40.722
MEDICINE-AESTHETICS	16.070
CELEBRITIES	25.555
SMOKE	41.040
ENTERTEINMENT	18.216

Table 5
Size, in tokens, after merging all the posts of a single class into a single document. These 10 documents were used to calculate the TF-IDF of the representations to see their domain relevance.

B. Distribution of Parts-of-Speech per Category of the extracted representation

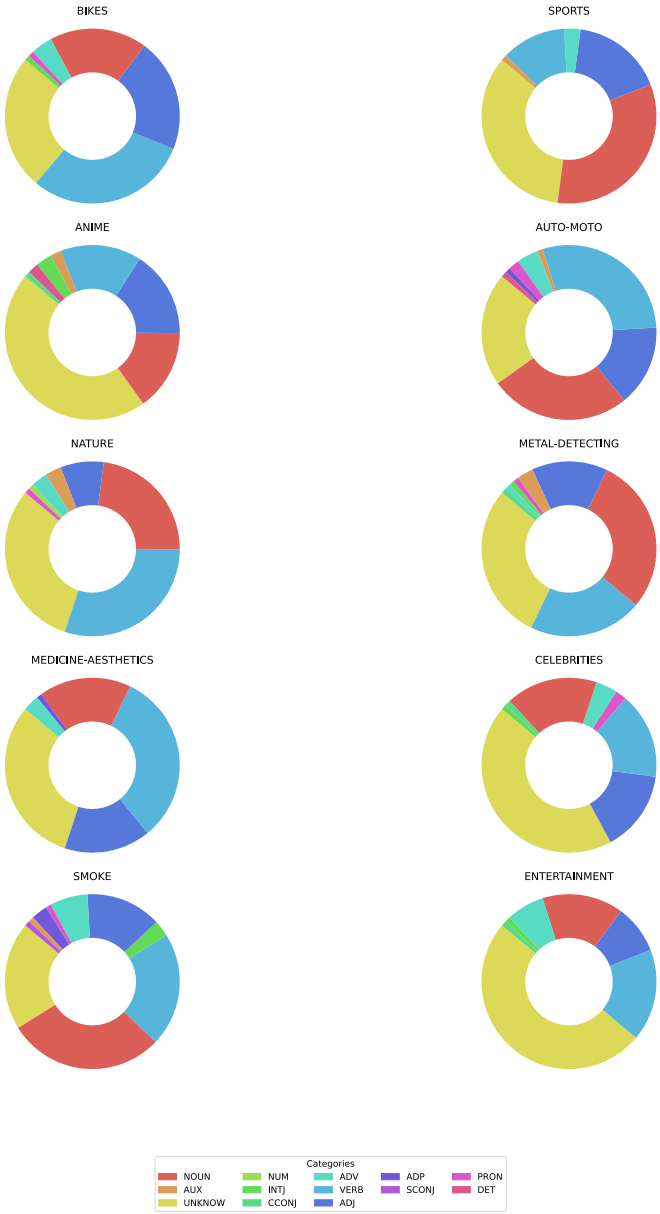

Figure 6: Distribution of Parts-of-Speech per Category of the representations extracted using the EOS method.