Lost in Labels: An Ongoing Quest to Optimize Text-to-Text Label Selection for Classification

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

[1]*Università di Pisa, Pisa, Italy*
[2]*TALIA s.r.l., Pisa, Italy*
[3]*ItaliaNLP Lab, CNR, Istituto di Linguistica Computazionale 'A.Zampolli', Pisa, Italy*

Abstract

In this paper, we present an evaluation of the influence of label selection on the performance of a Sequence-to-Sequence Transformer model in a classification task. Our study investigates whether the choice of words used to represent classification categories affects the model's performance, and if there exists a relationship between the model's performance and the selected words. To achieve this, we fine-tuned an Italian T5 model on topic classification using various labels. Our results indicate that the different label choices can significantly impact the model's performance. That being said, we did not find a clear answer on how these choices affect the model performances, highlighting the need for further research in optimizing label selection.

Keywords

encoder-decoder, label selection, topic classification

1. Introduction and Background

In recent years, the Sequence-to-Sequence paradigm has emerged as a highly popular approach in building cutting-edge Transformer-based Language Models [1, 2, 3]. This paradigm draws inspiration from earlier unified frameworks for Natural Language Processing (NLP) tasks [4, 5, 6], treating each task as a text-to-text transformation. In other words, it involves taking text as input and generating new text as output.

This unifying framework has proven to be a particularly effective transfer learning method, often outperforming previous models, e.g. BERT [7], in data-poor settings. Furthermore, the recent application and refinement of prompt-based tuning techniques for pre-trained Large Language Models (LLMs) have made this paradigm even more powerful, especially in few-shot and zero-shot learning scenarios [8].

In such a scenario, several studies have focused on defining methods for the formulation of prompts and the definition of *verbalizers*, i.e. mapping techniques between model-predicted words and task labels. As for the latter, the vast majority of studies have concentrated on devising automatic or semi-automatic approaches to create *verbalizers* that can be applied especially in zero- or few-shot configurations [9, 10, 11]. For instance, [12] proposed Petal, an approach for automatically finding the best words-label mapping by maximizing the likelihood of the training data. [13] instead developed ProtoVerb, a prototypical verbalizer that learns class prototypes from training data to build verbalizers automatically.

Nevertheless, few works have focused on investigating more deeply and systematically the effect that the choice of strings used to represent one (or more) labels has on model performance. Among these, [14] designed different label representations (e.g. canonical task labels, task-unrelated antonyms) and tested their impact with the T5 model on four classification tasks, showing that the performance was generally unaffected by the choice of label representation. Similarly, experimenting with the gender prediction task from the TAG-IT dataset [15], [16] noticed that while modifying the label representations did not affect the performance of the IT5 model [17], shuffling them for the topic classification task lead to worse results.

In this work, we present an evaluation of the impact of label selection on the performance of a Sequence-to-Sequence Model in a classification task. Specifically, we address the following research questions: i) Do the words used to represent the classification categories influence the model's performance? ii) Are there any relationship between classification categories and the words used to represent them that we can exploit to do label selection?

To investigate these questions, we conducted a series of experiments by fine-tuning the Italian version of the T5 model [17] on the topic classification task [15] using various labels. In particular, we defined different sets of labels and examined the model's performance for each of these sets. Additionally, we conducted an in-depth qualitative analysis to inspect which labels contribute

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, November 30 - December 2, 2023, Venice, IT

michele.papucci@talia.cloud (M. Papucci); alessio.miaschi@ilc.cnr.it (A. Miaschi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)

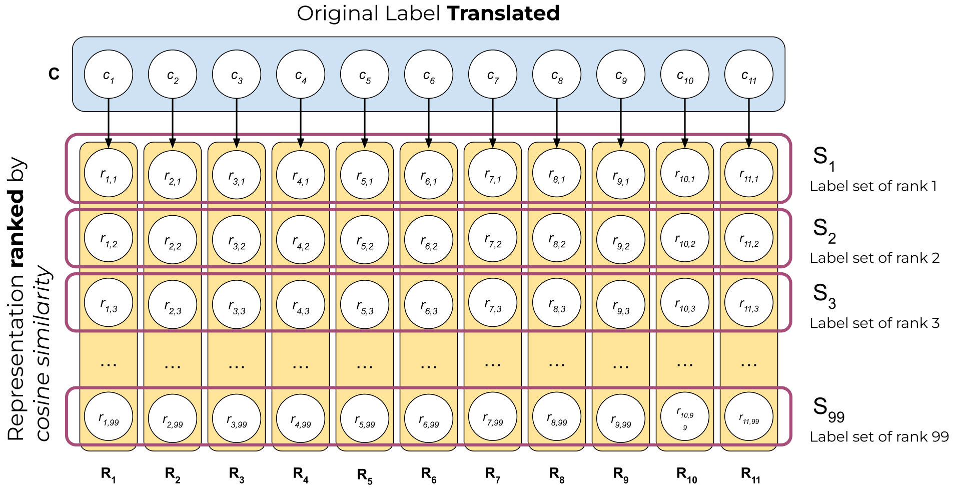
Original Label Translated
C
c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11
r1,1 r2,1 r3,1 r4,1 r5,1 r6,1 r7,1 r8,1 r9,1 r10,1 r11,1
r1,2 r2,2 r3,2 r4,2 r5,2 r6,2 r7,2 r8,2 r9,2 r10,2 r11,2
r1,3 r2,3 r3,3 r4,3 r5,3 r6,3 r7,3 r8,3 r9,3 r10,3 r11,3
...
r1,99 r2,99 r3,99 r4,99 r5,99 r6,99 r7,99 r8,99 r9,99 r10,99 r11,99
S1
Label set of rank 1
S2
Label set of rank 2
S3
Label set of rank 3
S99
Label set of rank 99
R1 R2 R3 R4 R5 R6 R7 R8 R9 R10 R11
Representation ranked by cosine similarity

Figure 1: The framework for the creation of the different sets of labels S_j ranked by cosine similarity.

most significantly to the improvement or decline in classification results and why that might be the case.

The remainder of the paper is organized as follows: in Sec. 2 we present our approach, introducing the data and the model we used (Sec. 2.1 and Sec. 2.2) and the experimental setting (Sec. 2.3). In Sec. 3 we discuss the obtained results and in Sec. 4 we conclude the paper.

Contributions. In this paper we: i) propose an evaluation of the influence that label selection has on the performance of a Text-to-Text Transformer model for classification; ii) investigate how the words used to represent the classification categories, in a multi-class classification task, impact task performance both globally, and at class-level; iii) investigate the existence of a relationship between classification categories and selected labels and how this connection can be leveraged to improve label selection.

2. Our Approach

In this section, we first define the data and the model used to perform our experiments. Then, we detail the experimental setting we devised to select the tested labels and fine-tune the T5 model.

2.1. 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 Italian and collected from different blogs. Each post is labelled with three different labels: age and gender of the writer and topic.

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.

In order to experiment with various possible combinations of labels, we have decided to focus only on the Topic classification task. Moreover, to have enough data to fine-tune the model, we decided to modify the original task as 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. Finally, since a fair amount of sentences were quite short, we decided to remove those shorter than 10 tokens. At the end of this process, we obtained a dataset consisting of 13,553 posts as training set and 5,055 posts as test set. The distribution of posts according to each label is reported in Table 1.

2.2. Model

We used the T5 base version pre-trained on the Italian language, i.e. IT5 [17][1]. 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.

2.3. Experimental Setting

As already introduced in Sec. 1, to investigate the influence of label selection on the model performance, we fine-tuned the IT5 model using different combinations of strings to represent the original classification categories. We will refer to the set of the original categories with C. We first translated the categories (as seen in Table 1) in Italian. (e.g. *Celebrities* into *celebrità*)[2]. Then, for each category c_i in C we created a set R_i composed by 100 string representations: 10 were selected from synonyms and related words to the original categories (including aforementioned translated ones), while the remaining 90 were randomly chosen from the most frequent nouns in the ItWac corpus [21]. Let $R_i = \{r_{i0}, r_{i1}, ..., r_{i99}\}$ be the set of labels for the category c_i, and r_{ij} be the j^{th} label in the set. Then, for each category c_i we ranked its corresponding set of labels R_i in descending order of similarity:

$$cs(c_i, r_{i0}) \geq cs(c_i, r_{i1}) \geq ... \geq cs(c_i, r_{i99})$$

Where $cs(c_i, r_{ij})$ is the cosine similarity between the average embedding of the subtokens of c_i and r_{ij}, extracted from the last encoding layer of the IT5 model.

Given the previously defined sets R_i, which contains the elements ranked by similarity, we created 100 sets of labels S_j (where j ranges from 0 to 99). Each set is defined as: $S_j = \{r_{0j}, r_{1j}, ..., r_{10j}\}$, where e.g. r_{0j} is the j^{th} ranked label for category c_0. As a consequence, S_0 contains the labels that achieved the highest cosine similarity with the original categories, while S_{99} is the set containing the lowest cosine similarities. An overview of our setting is shown in Figure 1.

We then fine-tuned IT5 for each ranked set of representation S_j. Each model was trained for 10 epochs and using f-score as the evaluation metric.

3. Results

Overall results Figure 2 summarizes the results obtained by the T5 models fine-tuned on the topic classification tasks according to the 100 different sets of labels (S_i).

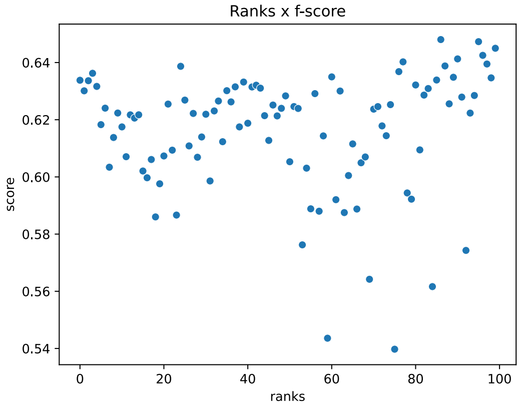

Figure 2: IT5 results (in terms of weighted f-scores) for each fine-tuning on the different sets of labels S_j.

At first glance, we can readily observe that the choice of words used to represent the classification categories has a considerable impact on the model's average performance. Indeed, we can see that the classification scores vary significantly, ranging from a minimum of 0.54 (rank 75) to a maximum of 0.65 (rank 86). Additionally, it is worth noting that the model trained with S_0, which contains the original translated labels, achieved an f-score of 0.63. This result indicates that simply using the original labels directly still provides a competitive performance. However, the significant fluctuations in the classification scores among the different sets S_j suggest that certain labels may still offer better performance than the original ones, while others may introduce noise or ambiguity, resulting in sub-optimal outcomes.

Interestingly, these findings appear to diverge from previous studies [14, 16], where the role of label representation was underestimated. While being a task-dependent issue, the role of label representation seems to have a large impact on model performance, especially for lower frequency labels, going as far as making certain labels range from being completely unpredictable to reaching satisfactory performances.

That being said, despite the differences in terms of weighted f-scores, there does not seem to be a clear correlation between the model's performance and the degree of "semantic" distance between the chosen labels and the original ones (represented by the rank j of the representation set). In fact, as the cosine similarity decreases between the selected representations and the original ones (from rank 0 to rank 99), there is no apparent trend in f-score values.

Per-label results In order to gain a more precise insight into the impact of the tested labels, Figure 3 illus-

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

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

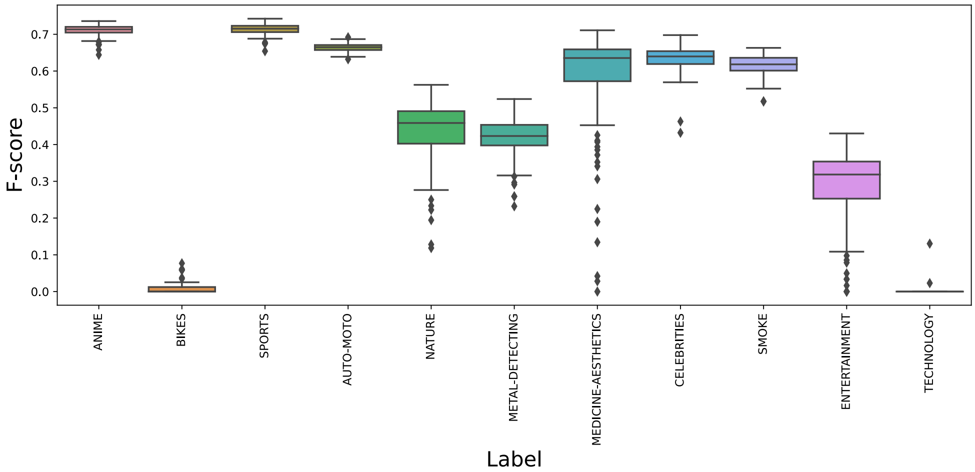

Figure 3: Boxplot showing the variation of the f-scores using different labels according to each classification category.

trates the variation of f-scores obtained with the 100 different sets of labels (S_i) for each individual category. Firstly, we can observe that the average results can vary significantly depending on the category under consideration. For instance, IT5 shows promising average performance in classifying posts related to *Anime*, *Sports* or *Auto-Moto*, while encountering difficulties in identifying posts annotated with the topics *Bikes* and *Technology*. This is possibly due to the fact that the posts belonging to the former categories are the most frequent in the entire dataset. Particularly noteworthy is the fact that, across almost all tested ranks, the model failed to correctly identify any posts related to *Technology*. This issue is likely attributed to the limited representation of this category within the dataset, further compounded by the original dataset configuration having more examples in the test set than in the training set (51 and 85 samples in the training and test sets respectively).

Analyzing the variation of results based on the labels used for representing the categories, we observe, in line with Figure 1, that the choice of the label often has a significant impact on the model's performance. While some labels exhibit relatively stable results with minor variations across different representations, such as *Anime*, *Bikes*, *Sports* and *Auto-Moto*, there are other instances where the selected labels lead to remarkable fluctuations in the model's performance. Notably, this behaviour emerges especially in the identification of posts related to *Nature*, *Metal-Detecting*, *Medicine-Aesthetics* and *Entertainment*. For these categories, IT5's classification performance can change drastically depending on the specific label. In some cases, the model manages to achieve quite good results, accurately classifying posts with a high f-score. However, in other instances, it struggles significantly, making erroneous classifications for the majority of cases. For instance, in the case of *Medicine-Aesthetics*, the f-score reaches a maximum of 0.71 when the label is represented by the term *acuto* but it fails to correctly classify any instance (f-score = 0) when the label is represented as *proprio*. This highlights how the choice of the label can significantly impact IT5's classification performance across different topics and therefore, suggests the importance of exploring optimized selection strategies to maximize the model performance.

To obtain a more comprehensive qualitative perspective of these findings, we include in Figure 4 the top and bottom 10 representations that maximized/minimized the f-score values for the four aforementioned categories. As we can observe, among the four considered categories, only one (*Medicine-Aesthetics*) contains the original label, i.e. the one with cosine similarity equal to 1 (*medicina*), in the top 10 representations. For the other categories, the absence of the original label seems to suggest that the chosen word for the label, which should be the closest one to the reference topic, may not be the one that can maximize the results. When analyzing individual words, it becomes evident that not all words contributing to the model's best performance belong exclusively to the domain of the considered category. Surprisingly, words such as *cinema* and *sitcom*, seemingly related to the *Entertainment* domain, are among those that most negatively impact the model's f-scores. Nevertheless, *Medicine-Aesthetics* shows an exception, with several words aligned with the category's domain, e.g. *benessere*, *medicina*, *dottoressa* e *sensibilità*. Lastly, it is worth noticing that the performance drop is mostly label-dependent, and there is a significant

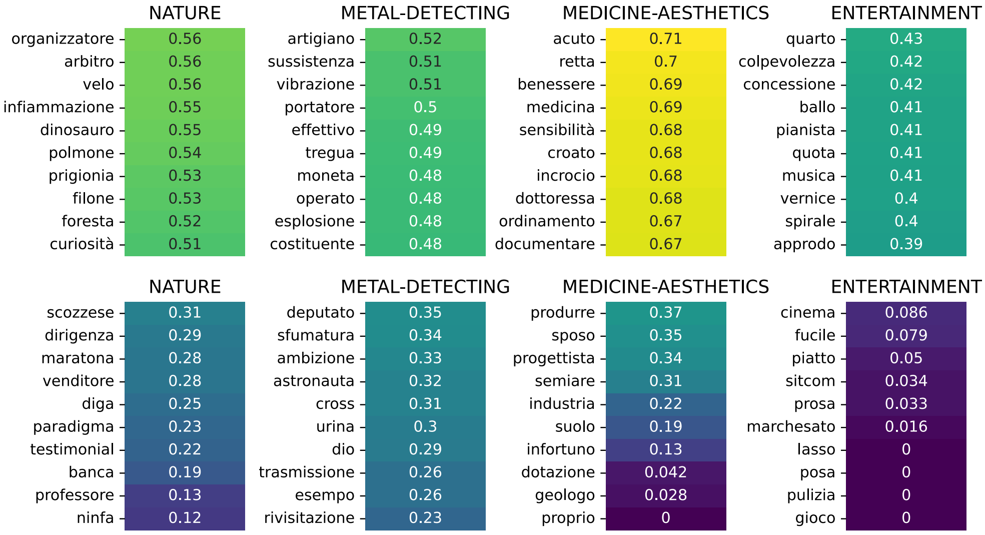

Figure 4: Top and bottom 10 labels that maximize/minimize the results for the most varying categories (*Nature*, *Metal-Detecting*, *Medicine-Aesthetics* and *Entertainment*).

Categories	Spearman	p-value
Entertainment	0.29	0.003 *
Auto-Moto	0.05	0.62
Medicine-Aesthetics	-0.02	0.85
Bikes	-0.05	0.61
Anime	-0.10	0.37
Technology	-0.12	0.21
Smoke	-0.20	0.04 *
Sports	-0.22	0.03 *
Nature	-0.25	0.01 *
Metal-Detecting	-0.35	0.00 *
Celebrities	-0.45	0.00 *

Table 2
Spearman correlations between f-scores and label similarities (cosine similarity) for each category. Statistically significant correlations are marked with *.

difference between the most- and least-performing representations for the four categories. In fact, while *Nature* and *Metal-Dectecting* exhibit a relatively modest decrease (around .20 f-score points), *Medicine-Aesthetics* and *Entertainment* display a far more pronounced difference in performance.

3.1. Correlating Model Performance and Tested Representations

Having analyzed the model's performance and assessed the impact of words used to represent the categories on the classification results, we decided to explore the existence of any relationship between the model's performance and the employed words.

Semantic Similarity Initially, we aimed to ascertain whether there is a correlation between the words that are more/less semantically similar to the original categories and the performance of IT5. To achieve this, we computed the Spearman correlation between the T5 model's performance and the cosine similarity values calculated to construct the 100 sets for each label S_j. The results of these correlations are presented in Table 2[3]. As observed, 6 out of the 11 classification categories exhibit statistically significant correlations. Among these, only one correlation is positive (*Entertainment*), while the others show negative correlation values. This outcome is quite unexpected as it seemingly implies that the improvement in the model's performance is linked to a decrease in semantic similarity. However, it is crucial to emphasize that the correlation values are not particularly high, and thus, we cannot draw any conclusion about these results. Moreover, it is important to consider that while cosine similarity can serve as a useful measure of similarity between embeddings, it may not encompass the entire semantic space.

Internal Similarity Since the similarity between selected labels' within each set could potentially impact the model's performance, we conducted an additional test to investigate whether higher semantic similarity among

[3] In Appendix A we also reported the scatterplots showing the relationship between f-scores and cosine similarity values for these labels.

representations within a set could negatively affect the performance of IT5. To achieve this, we computed the *"inner similarity"* of each set, defined as the average cosine similarity of all possible distinct label combinations[4]. Subsequently, we computed the Spearman correlation between each set's *"inner similarity"* and the f-scores obtained by the model fine-tuned with it. Although the values of *"inner similarities"* vary considerably across the sets (ranging from a similarity of 0.69 for rank 0 to 0.38 for rank 100), we did not find a statistically significant correlation with the model's performance (Spearman = 0.01, p-value = 0.90). These results suggest that, despite the sets exhibited considerable variation in terms of inner similarity, the similarity between the representation didn't plainly affect the model's performance.

Representations Frequencies Finally, since the aforementioned results have demonstrated that different labels have an impact on the model's performance, we decided to investigate whether this impact could be somehow related to the frequency of these representations within the model's training dataset. To this end, we computed the absolute frequency of each label used in our experiments (11 labels per 100 sets, totalling 1100 words) within the Italian version of the mC4 Corpus, i.e. the corpus on which IT5 was trained. Subsequently, we calculated the correlation between the scores obtained by IT5 for each label of each set R_i and the corresponding frequencies of each label found in the mC4 corpus. Among the 11 categories present in the dataset, only one showed a statistically significant correlation, *Smoke*, with a Spearman correlation value of -0.25[5]. This result suggests that, at least for this particular category, a decrease in the label's frequency in the training corpus corresponds to an increase in the model's performance. However, the fact that only one representation exhibits a significant correlation and that this correlation is not particularly high once again prevents us from drawing any conclusive findings. Thus, it underscores the need to explore other strategies in the future for label selection.

4. Conclusion

In this work, we presented an evaluation of the impact of label selection on the performance of a Sequence-to-Sequence Model in a classification task. By fine-tuning the Italian version of the T5 model on a topic classification task, we explored various sets of labels and examined their influence on the model's performance.

Our results indicate that the choice of words used to represent the classification categories can have a significant impact on the model's performance. While some labels led to competitive results, others resulted in suboptimal outcomes, with noteworthy variations in the classification scores. This finding diverges from previous studies that suggested label representations had little impact on model performance.

Interestingly, the correlation between the model's performance and the degree of "semantic" distance between the chosen labels and the original ones was not clear. While some labels exhibited statistically significant correlations, they were either positive or negative, indicating that higher or lower semantic similarity did not consistently lead to better performance.

In conclusion, our findings suggest that the choice of the label is not a trivial matter and can have a significant impact on the performance of Sequence-to-Sequence Models in classification tasks. To maximize performance, it is essential to explore optimized label selection techniques that are carefully selected and tailored to the specific task and dataset.

Future research could focus on developing more sophisticated methods for label selection, taking into account not only semantic similarity but also other relevant factors. Additionally, it would be valuable to investigate the generalizability of these findings across other languages and models, and in order to gain a more comprehensive understanding of the influence of label selection on different NLP tasks.

Acknowledgments

This work has been supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] 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.

[2] 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.

[3] 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.

[4] B. McCann, N. S. Keskar, C. Xiong, R. Socher, The natural language decathlon: Multitask learn-

[4] As defined in Sec. 2.3, a label is represented as the average embedding of each subtoken in the string.

[5] The table with all the correlations is reported in Appendix B.

ing as question answering, arXiv preprint arXiv:1806.08730 (2018).

[5] N. S. Keskar, B. McCann, C. Xiong, R. Socher, Unifying question answering, text classification, and regression via span extraction, arXiv preprint arXiv:1904.09286 (2019).

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners (2019).

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[8] 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).

[9] C. Song, F. Cai, J. Zheng, W. Chen, Z. Pan, Metric sentiment learning for label representation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1703–1712. URL: https://doi.org/10.1145/3459637.3482369. doi:10.1145/3459637.3482369.

[10] W. Jiang, Y. Zhang, J. Kwok, Effective structured prompting by meta-learning and representative verbalizer, in: International Conference on Machine Learning, PMLR, 2023, pp. 15186–15199.

[11] K. Ji, Y. Lian, J. Gao, B. Wang, Hierarchical verbalizer for few-shot hierarchical text classification, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2918–2933. URL: https://aclanthology.org/2023.acl-long.164.

[12] T. Schick, H. Schmid, H. Schütze, Automatically identifying words that can serve as labels for few-shot text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5569–5578. URL: https://aclanthology.org/2020.coling-main.488. doi:10.18653/v1/2020.coling-main.488.

[13] G. Cui, S. Hu, N. Ding, L. Huang, Z. Liu, Prototypical verbalizer for prompt-based few-shot tuning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7014–7024. URL: https://aclanthology.org/2022.acl-long.483. doi:10.18653/v1/2022.acl-long.483.

[14] 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, Association for Computational Linguistics, Suzhou, China, 2020, pp. 160–164. URL: https://aclanthology.org/2020.aacl-srw.23.

[15] A. Cimino, F. Dell'Orletta, M. Nissim, Tag-it@ evalita 2020: Overview of the topic, age, and gender prediction task for italian, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).

[16] 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.

[17] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[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] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

A. Appendix A

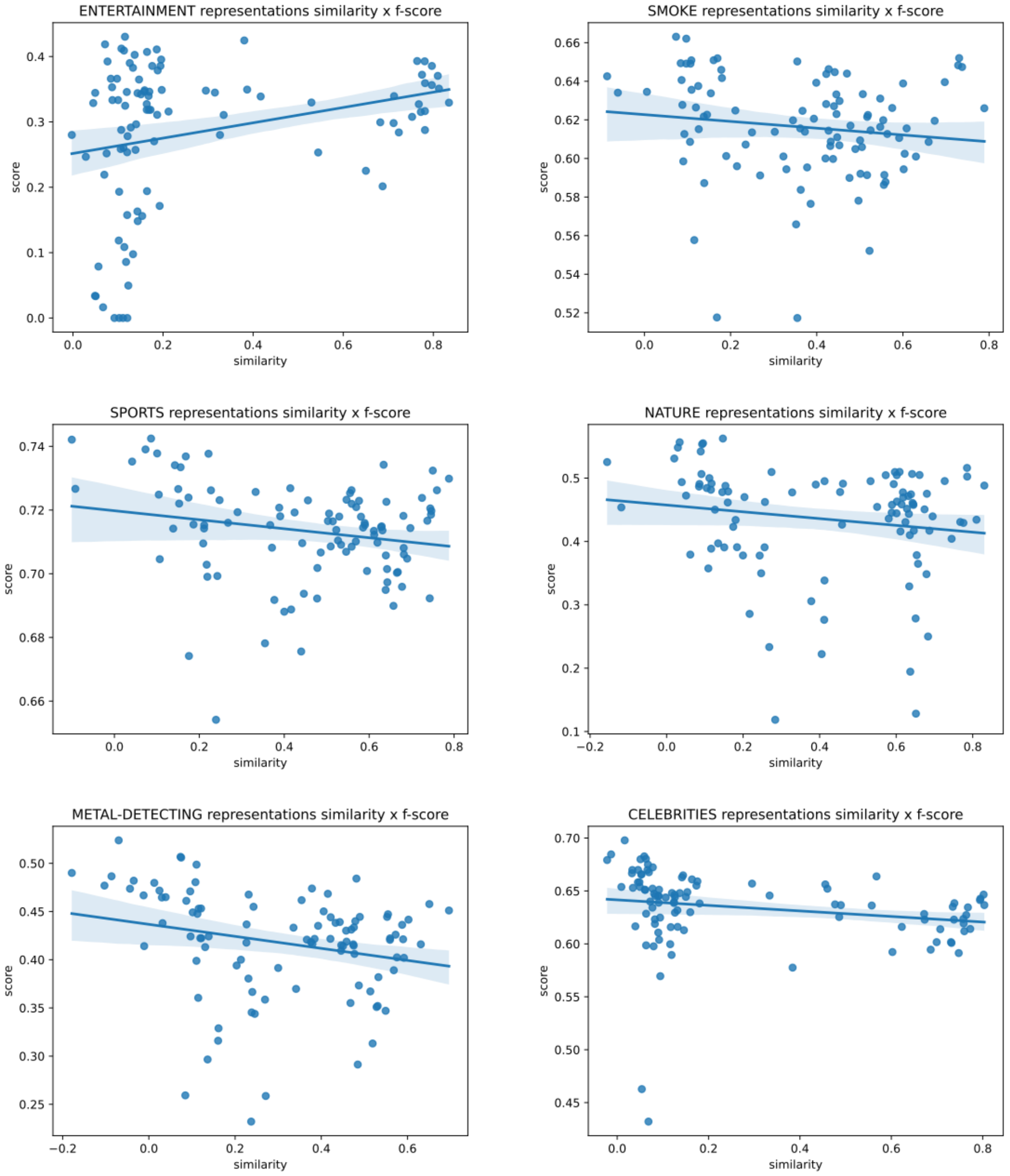

Figure 5: Scatterplot showing the relationship between f-scores and cosine similarity values for the 6 categories that exhibited a statistically significant correlation.

B. Appendix B

Categories	Spearman	p-value
Medicine-Aesthetics	0.13	0.20
Nature	0.06	0.54
Sports	0.04	0.66
Bikes	0.01	0.94
Technology	-0.02	0.88
Anime	-0.02	0.84
Entertainment	-0.03	0.75
Auto-Moto	-0.05	0.62
Metal-Detecting	-0.06	0.57
Celebrities	-0.06	0.54
Smoke	-0.25	0.01 *

Table 3
Spearman correlations between f-scores and labels absolute frequencies (computed in the Italian mC4 Corpus) for each category. Statistically significant correlations are marked with *.