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

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

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

talia

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Open Issue: Few works have investigated the importance of the string representation of classes in T2T classification tasks.

Introduction

talia

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Open Issue: Few works have investigated the importance of the string representation of classes in T2T classification tasks.

Our Contribution: We present an investigation on the <u>importance of string representations for model performances</u>, and on the relationship between the classes and the strings that represent them.

Introduction

Dataset dimension: 13.553 posts formed the training dataset and 5055 the test dataset.

Task: We focused on the Topic Classification Task, which is a multilabel classification tasks with **eleven classes**.

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

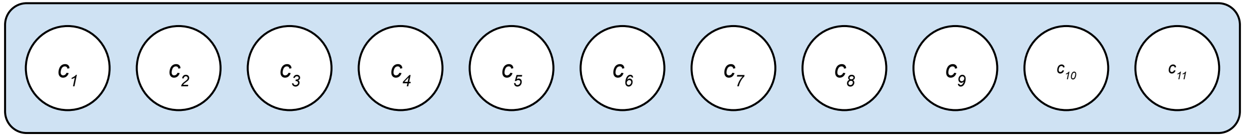
TAG-IT Dataset (Cimino et al., 2020)

Model: IT5 (Sarti e Nissim, 2022), an encoder-decoder architecture based on T5 and trained on the italian sentences cleaned and retrieved from the mC4 corpus.

Idea: Finding a meaningful relationship between the classes and their string representation to do **label representation selection**. In particular, we tried to see whether the *cosine distances* between the **class** and **its string representation** are correlated to the model's performance on that class.

Original Classes **Translated**

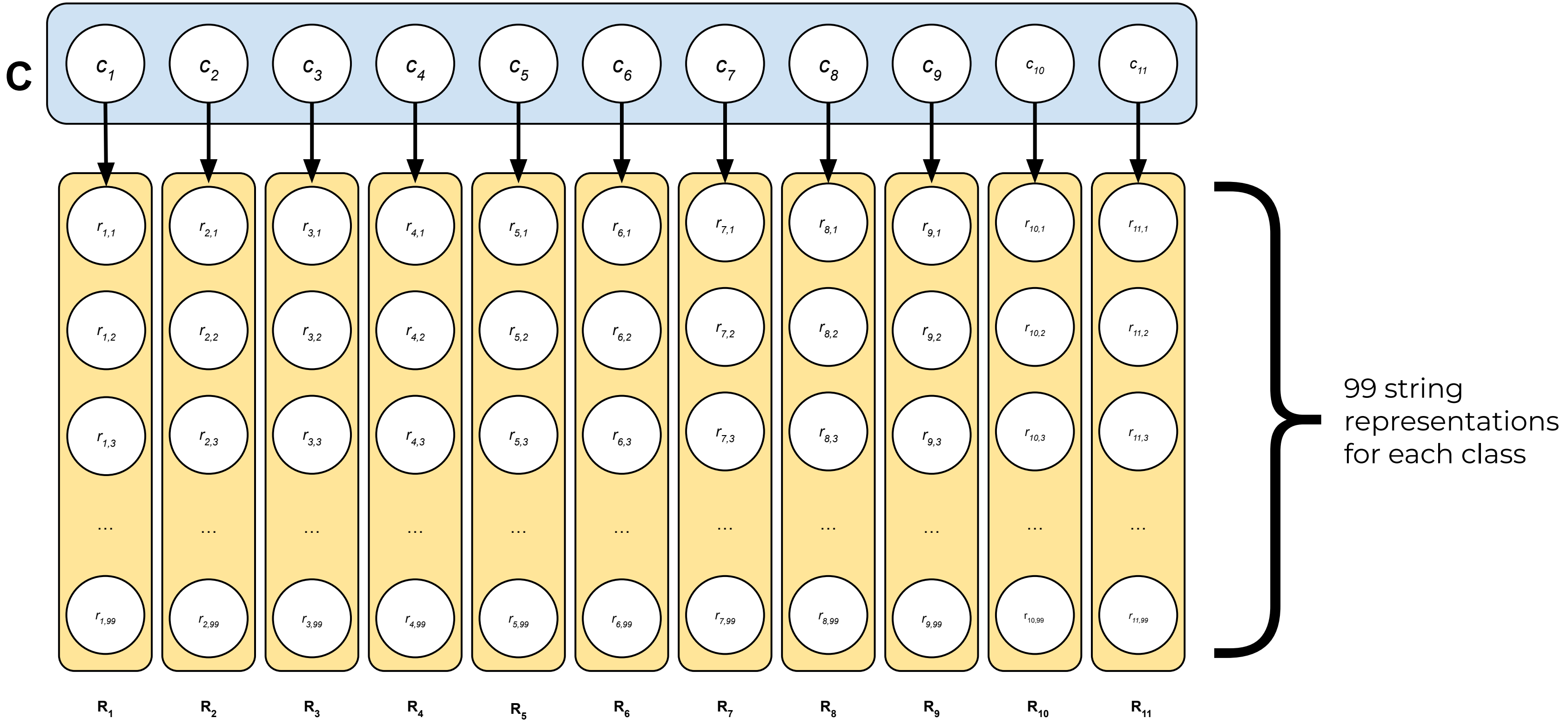
C



c_1 = **Anime** ⟶ *Anime*
c_2 = **Auto-Moto** ⟶ *Automobilismo*
c_3 = **Bikes** ⟶ *Bicicletta*
c_4 = **Celebrities** ⟶ *Celebrità*
c_5 = **Entertainment** ⟶ *Intrattenimento*
c_6 = **Medicine-Aesthetics** ⟶ *Medicina*
c_7 = **Metal-Detecting** ⟶ *Metal Detector*
c_8 = **Nature** ⟶ *Natura*
c_9 = **Smoke** ⟶ *Fumo*
c_{10} = **Sports** ⟶ *Sport*
c_{11} = **Technology** ⟶ *Tecnologia*

These representations of the original classes have been used to calculate the *cosine similarities* between the **classes** and the **candidate representations.**

Original Classes **Translated**

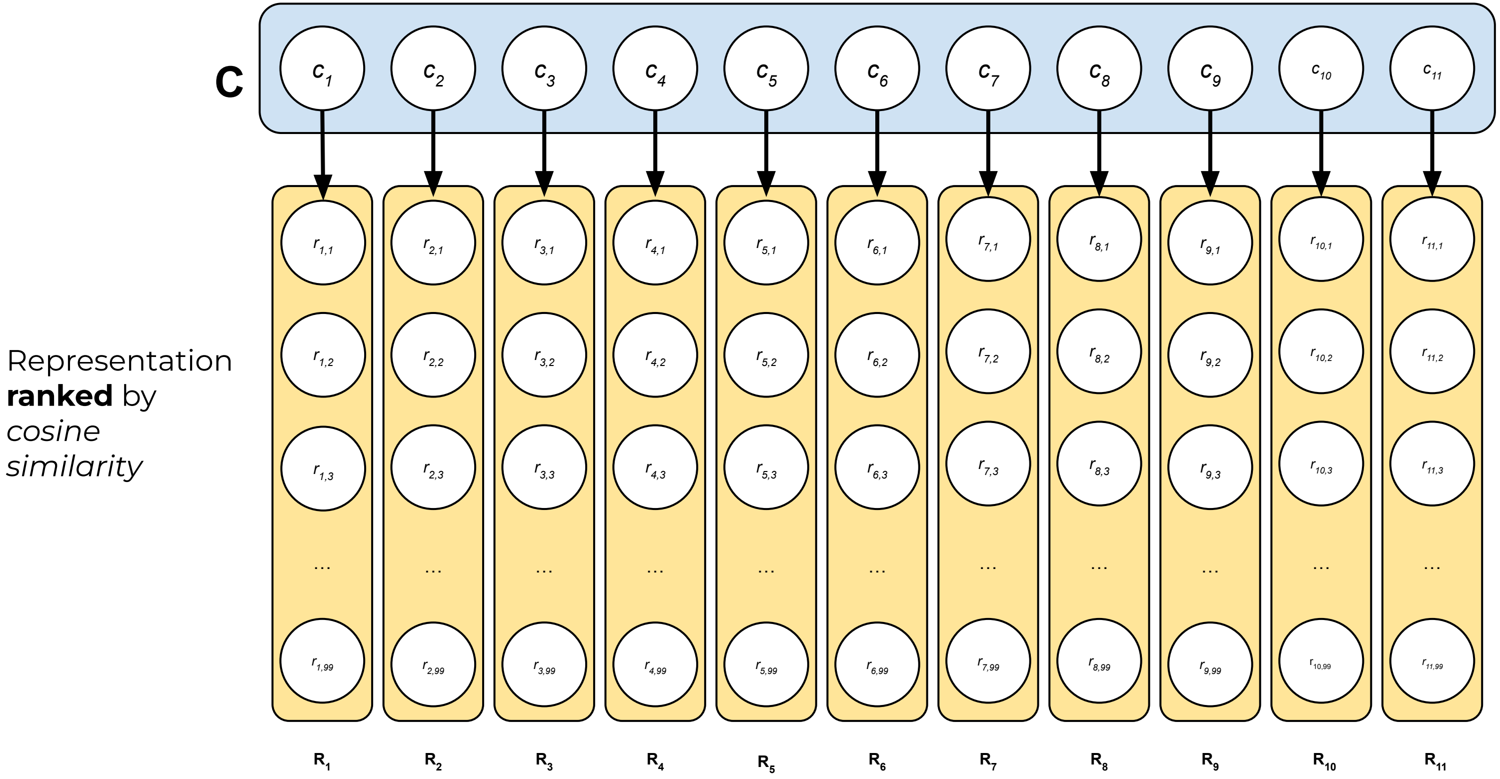


We have 100 representations: the original class translated + 10 manually selected from the class **synonyms** + 90 **randomically chosen noun** between the most frequent in the ItWac Corpus (Baroni et al., 2009)

Experimental Setting

talia

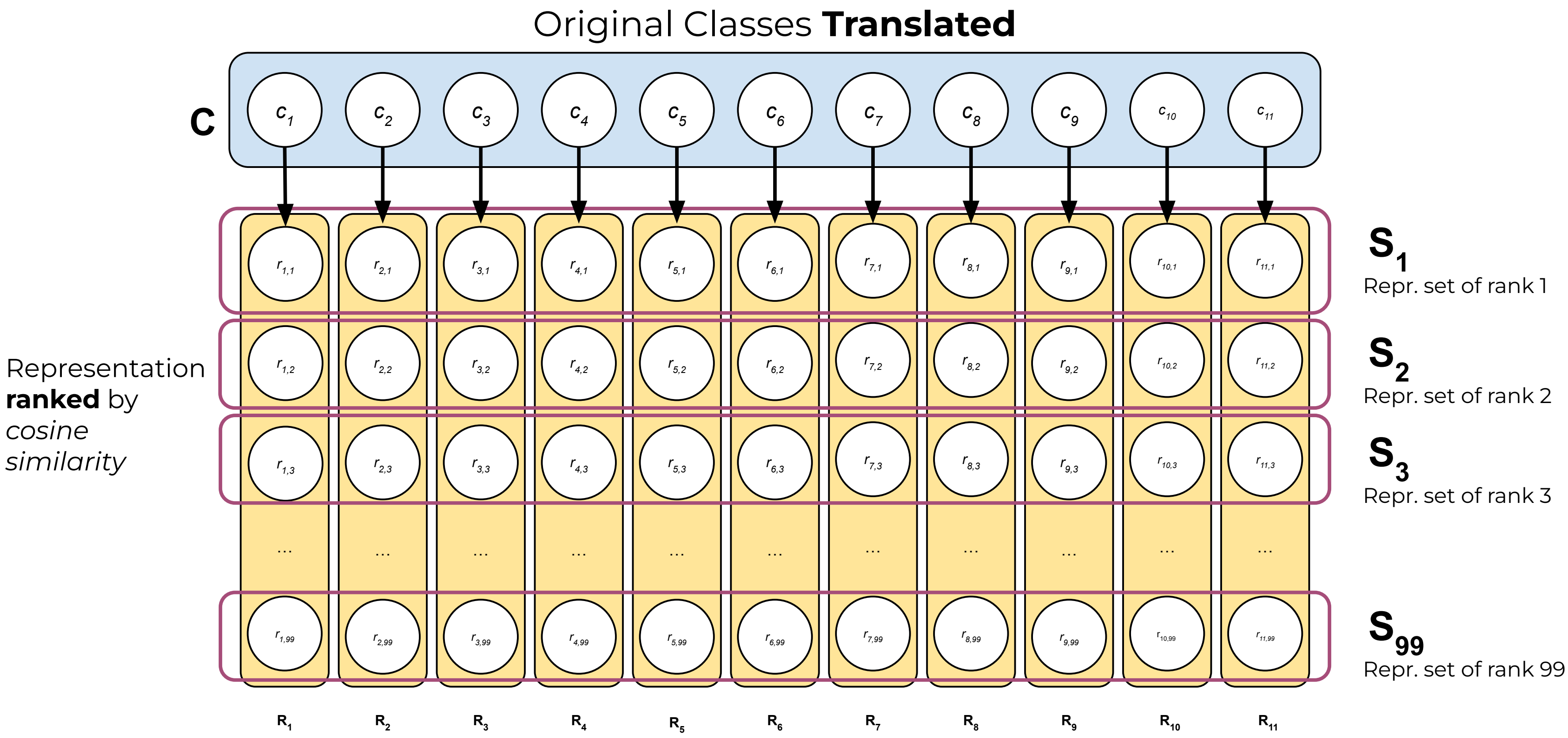
Original Classes **Translated**



The 100 representation have been ranked by the *cosine similarity* between the original class translated and the candidate representation. The similarity is computed between the **IT5 embedding vectors** that represent the strings

Experimental Setting

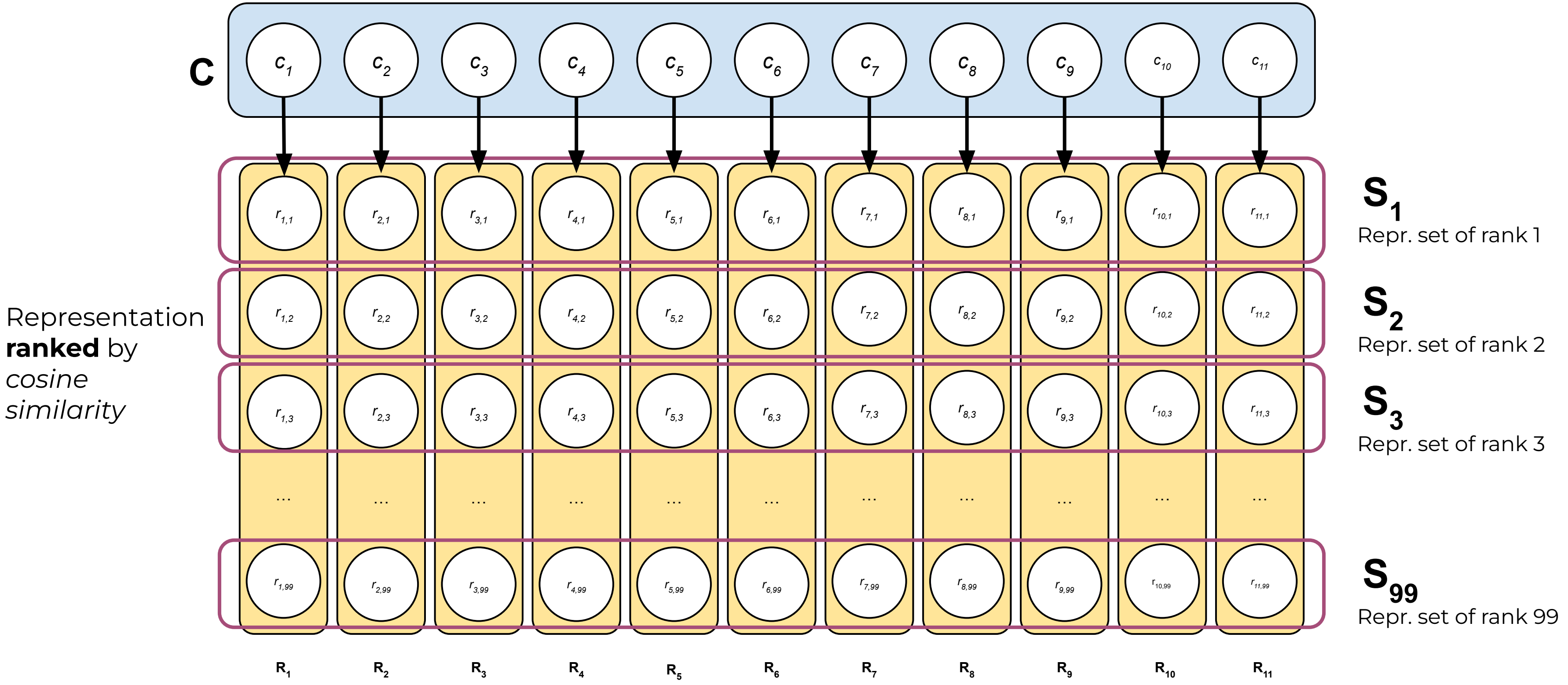
talia



Experimental Setting

talia

Original Classes **Translated**

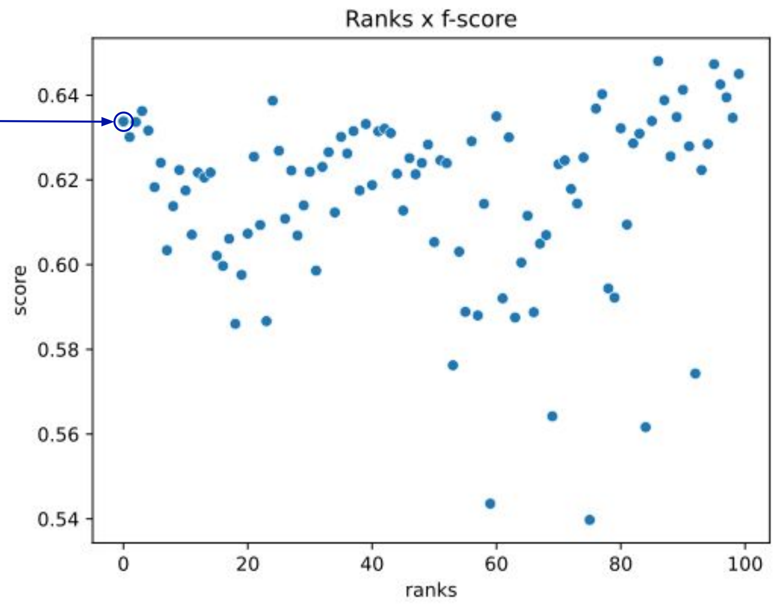


Fine-Tuning: After creating the 100 sets of representation S_1, …, S_{99} we fine-tuned 100 IT5 models on each of this sets of label representations.

Experimental Setting

talia

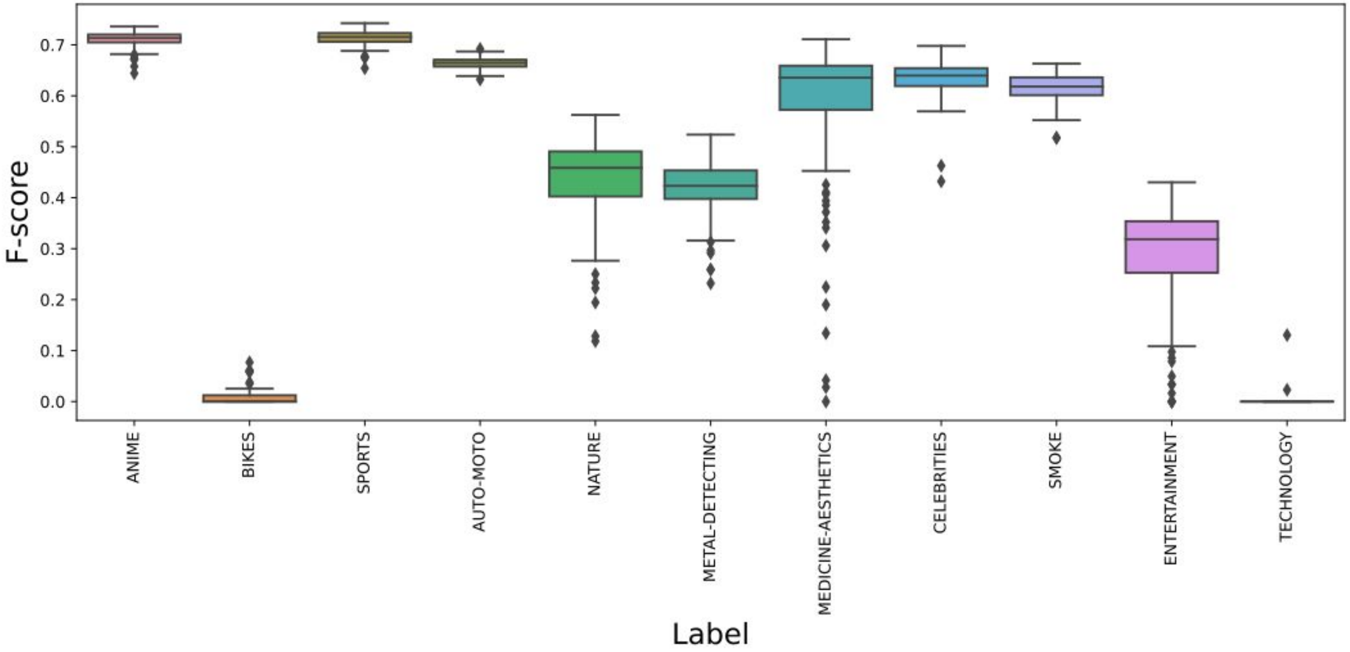
The model fine-tuned with the **translated class names** performs competitively but it's far from the best.



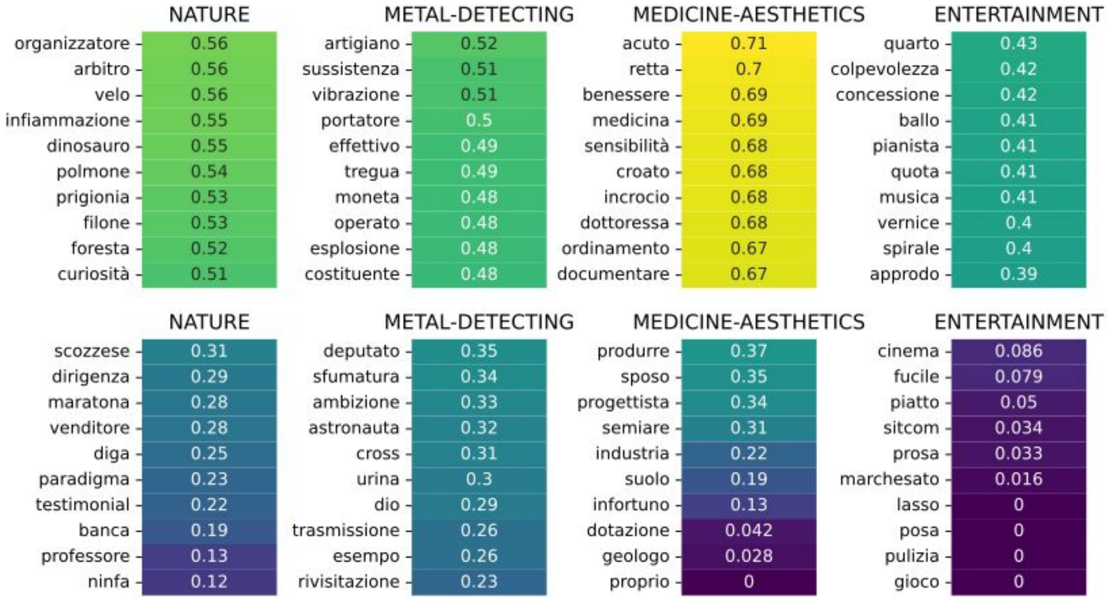
Overall results: we can observe how the choice of string representation has a considerable impact on the models performances. However, no correlation was found between the **representation' ranks** and the **model weighted F-score**.

Results

talia



Per-class analysis: The way the classes are represented is especially important for lower-frequency classes where the f-score variations are greater. In some of the least frequent classes the f-score ranges from zero to acceptable performances.



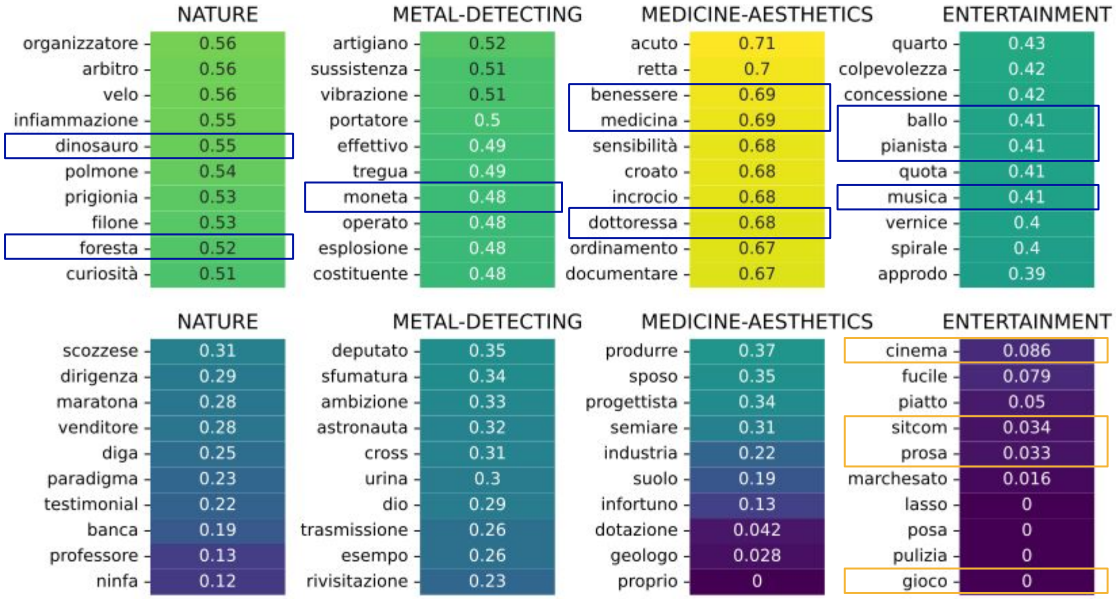
Top 10 representation by f-score

Worst 10 representation by f-score

Per-class analysis: looking at the classes with the **highest f-score variance** <u>there is no clear indication to which representations work better.</u>

Results

talia



Top 10 representation by f-score

Worst 10 representation by f-score

Per-class analysis: looking at the classes with the **highest f-score variance** there is no clear indication to which representations work better. <u>The placement of in-domain words in the f-score ranking doesn't indicate that those words work better.</u>

Results

talia

Internal similarity: we calculated Spearman correlation between the f-score and the *internal similarity score* of a set.

The *internal similarity score* of a set S was defined as the **average cosine distance between all possible distinct combination of representation couples in a set**.

The score varied considerably between sets, however, when we calculated Spearman between:

- the **100 internal similarity scores** (one for each representations sets);
- the **100 f-score** (one for each model fine-tuned on the representations set);

We found a **Spearman of 0.01** with a **p-value of 0.9**, indicating no apparent correlation between how semantically similar are the representation between themselves in a set, and how the model performs.

Representation frequencies: we calculated the Spearman correlation per-class between the **f-scores** and the **absolute frequency** of the representation in the mC4 training corpus of IT5.

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 *

The only statistically significant results was on Smoke.

Results - Label Selection Strategies (2)

talia

Representations are important: our results indicate that for tasks such as Topic classification where lexical information are important, the choice of label representation is critical to model performance, especially for low-frequency classes where the classification f-score can vary from 0 to competitive results, but we didn’t find out why.

Representations are important: our results indicate that for tasks such as Topic classification where lexical information are important, the choice of label representation is critical to model performance, especially for low-frequency classes where the classification f-score can vary from 0 to competitive results, but we didn't find out why.

Representation similarity is not that important: we found that how similar the representation is to the original class doesn't seem to affect the classification results. This, however, could be attributed to poor choice of the initial class name for the category or *cosine similarity* not being an effective measure of semantic similarity for this purpose.

Conclusion

Representations are important: our results indicate that for tasks such as Topic classification where lexical information are important, the choice of label representation is critical to model performance, especially for low-frequency classes where the classification f-score can vary from 0 to competitive results, but we didn’t find out why.

Representation similarity is not that important: we found that how similar the representation is to the original class doesn’t seem to affect the classification results. This, however, could be attributed to poor choice of the initial class name for the category or *cosine similarity* not being an effective measure of semantic similarity for this purpose.

Representation choice is not trivial: finding class representation in T2T classification scenarios **is not trivial** and we couldn’t find a simple and effective way to choose them in a way that maximise the performances. <u>We propose to call the task of finding the best class representation in a T2T classification scenario **Automatic Class Label Selection**</u>, and future research should focus on developing an effective way to solve it.

Conclusion

talia

Thank you for your attention!

Contacts

Mail: michele.papucci@talia.cloud
Twitter: @mpapucci_
Linkedin: linkedin.com/in/michelepapucci

talia

Semantic Similarity: we calculated Spearman correlation per-class between the **f-score** of a model trained with a specific representation, and the representation's ***cosine similarity*** with its class.

We found 6 had statistically significant correlation, with most of the correlation being negative, implying that to a **higher similarity** between the representation and the class name **corresponds a lower f-score**.

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 *

Results - Label Selection Strategies (3)

talia