Improving Keyword-Based Topic Classification in Cancer Patient Forums with Multilingual Transformers

T.M. Buonocore[a], E. Parimbelli[a], L. Sacchi[a], R. Bellazzi[a], L. del Campo[b], S. Quaglini[a]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy,
[b] AIMAC, Italian Association of Cancer patients, relatives and friends, Rome, Italy

Abstract

Online forums play an important role in connecting people who have crossed paths with cancer. These communities create networks of mutual support that cover different cancer-related topics, containing an extensive amount of heterogeneous information that can be mined to get useful insights. This work presents a case study where users' posts from an Italian cancer patient community have been classified combining both count-based and prediction-based representations to identify discussion topics, with the aim of improving message reviewing and filtering. We demonstrate that pairing simple bag-of-words representations based on keywords matching with pre-trained contextual embeddings significantly improves the overall quality of the predictions and allows the model to handle ambiguities and misspellings. By using non-English real-world data, we also investigated the reusability of pretrained multilingual models like BERT in lower data regimes like many local medical institutions.

Keywords:

Natural Language Processing, Classification, Community Health Services.

Introduction

Despite the unrestrainable growth of social media and microblogging in communication, online forums still play an important role in connecting people whose lives have been touched by cancer. These communities - made by patients, relatives, friends, and trained volunteers - create networks of mutual support on different cancer-related topics, covering different needs people develop while facing the various stages of the disease. On these medical message boards (MMBs), users communicate asynchronously sharing details about their experiences and answering questions about symptoms, treatments, diagnosis, etc. MMB discussions include posts from authors with different backgrounds, expertise, education and writing skills, representing an extensive source of heterogeneous textual information [11]. Due to its particularly unstructured and noisy nature, this source of information is often underutilized and can be mined to offer better assistance in many applications [1,12]. This work leverages MMB posts focusing on topic classification for posts filtering, forum moderation or thread recommendations for new discussions.

Text classification (including topic classification) refers to the Natural Language Processing (NLP) task of assigning a sentence or document an appropriate category by learning associations and patterns between pieces of the text. In order to be understood by machine learning models, documents, i.e., posts, must be transformed into a numerical representation before being classified.

Count-based approaches like Bag-of-Words (BoW) represent posts based on word distributions that are estimated by counting the frequency of occurrence of each word of the vocabulary in a post. These approaches are intuitive, interpretable, and easy to implement, although they do not consider the word order nor the syntactic structure and are less suitable for short texts due to their sparsity and high dimensionality [16]. Some of the BoW limitations have been addressed in the past by applying dimensionality reduction techniques [5] or by counting over a limited set of informative concepts or keywords instead of the whole vocabulary [6,16].

On the other hand, prediction-based approaches natively learn dense continuous vectors by solving a prediction task rather than counting, providing representations based on the model's weights where the semantic similarities between the words are well-preserved [9]. One of the most advanced ways to produce these vectors, commonly referred to as embeddings, is BERT, a language representation model based on Transformers that also addresses Italian in its multilingual version [4].

CAPABLE[1] is a European-funded project aimed at developing a decision support system to improve the quality of life of cancer patients treated at home, by combining technologies for data and knowledge management with socio-psychological models and theories [2]. As a partner of the CAPABLE project, the Italian Association of Cancer patients, relatives, and friends (Aimac)[2] is making the data collected in its discussion forum available for several analytics tasks, including the identification of patient needs. This forum offers a virtual place where people facing directly or indirectly cancer can meet, share their experiences and discuss [7].

In this paper, we exploit the Aimac forum as a source of information for topic classification to:

1. Investigate the added value of combining simple keywords matching with advanced prediction-based representations for topic classification.
2. Propose a model for MMB moderation support based on Multilingual BERT, following an automated approach for training data annotation and text representation.
3. Test the effectiveness of multilingual pre-trained models on Italian MMBs.

Furthermore, as most of the advancements in NLP are geared towards the English language [8], we believe that investigating transfer learning with pre-trained multilingual Transformers using limited real-world data may help to reduce the need for heavily engineered architectures and massive amounts of training data, requirements that cannot be matched in many

[1] CAPABLE website: https://capable-project.eu
[2] Aimac website: https://www.aimac.it

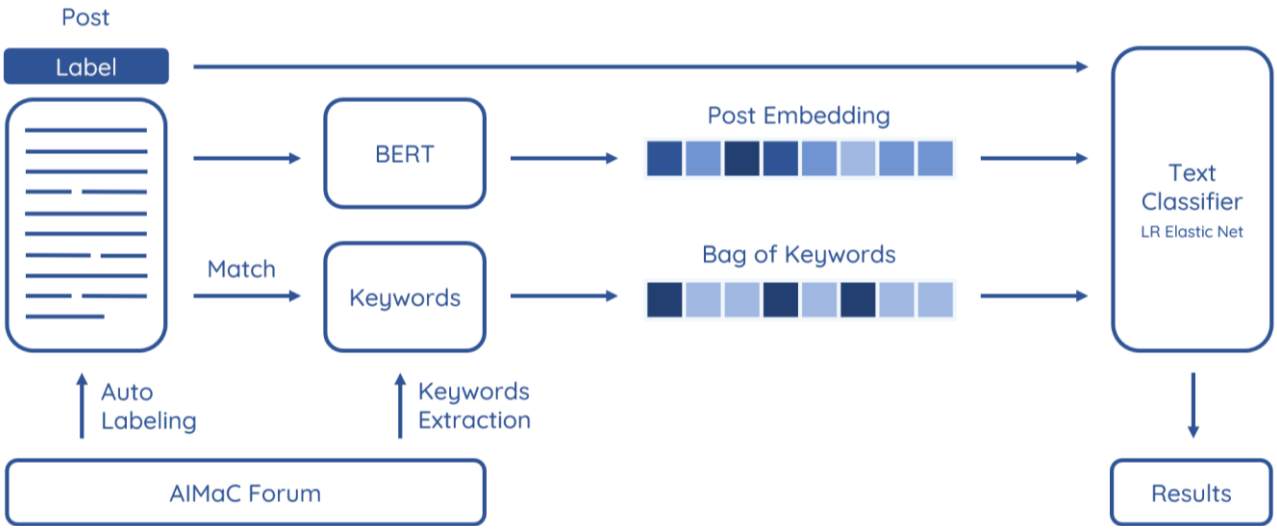

Figure 1 – Topic classification pipeline for Aimac forum posts

local medical centers of non-English-speaking countries that want to develop clinical applications based on Natural Language Processing techniques.

Methods

The Aimac forum includes 74930 posts, written in Italian by 3955 unique users. Discussions are distributed in 35 subforums based on their main subject. Each of these subforums hosts multiple threads. The average number of threads in a forum is 117, while the average number of posts is 2140, varying from 6 to 35095. In this paper, we consider the subforums with the highest participation in terms of opened threads and published posts, excluding the general-purpose ones (e.g., "Introduce Yourself", "Scattered Thoughts", "Staff News"). This leaves us with 8 hot-topics: "Prostatic Cancer", "Pancreatic Cancer", "Lung Cancer", "Colorectal Cancer", "Head and Neck Cancer", "Brain Tumor", "Breast Cancer" and "Liver Cancer".

Data Preparation and Pre-processing

Forum posts have been filtered considering only the first *M* messages of each thread. Each post has been truncated to the first *W* tokens and annotated leveraging the inner structure of the forum itself with a One-vs-Rest approach: positive (*on-topic*) if the message belongs to the subforum of interest, negative (*off-topic*) otherwise. The *M* and *W* parameters are manually defined, and different configurations have been tested.

Punctuation and special tokens (e.g. HTML tags, escape sequences, emoticons) have been removed from the text. No stemming nor lemmatization has been deployed, as well as stop words removal. We applied such minimal preprocessing because non-trivial words may still provide useful contextual information for Transformer-based models like BERT [14].

Binary labeling of posts introduces class imbalance, which has been addressed with undersampling. We split the data randomly with stratification into an 80% training set and a 20% test set, using 20% of the training set as the validation set for BERT fine-tuning.

Text Representation

In order to perform text classification, the first major step is to define a mapping strategy to represent posts as numerical vectors. Our work explores the use of two types of representation, as shown in Figure 1, where we illustrate the pipeline described in this paper.

Bag of Keywords

Keywords matching can be used as a simple and concise way to represent text [6]. Given a set of keywords *K*, each post *P* is mapped in a Bag-of-Keywords (BoK) vector *p* according to Equation 1.

$$P \rightarrow p = [p_1, \dots, p_K], \qquad p_i = \begin{cases} 1 & k_i \in P \\ 0 & k_i \notin P \end{cases} \qquad Eq.1$$

Despite ignoring context and word order, BoW approaches constitute a sufficient description of the text content in many applications [3]. The sets of keywords have been automatically extracted with TF-IDF on subforums keeping only the top *K* terms for each topic, stop words excluded. *K* is a user-defined hyperparameter. Results for different values of *K* are reported in the Results section.

Figure 2 – Italian keyword clouds for Pancreatic Cancer (top) and Prostatic Cancer (bottom) when K=100.

Figure 2 shows an example of an automatically extracted keywords set for pancreatic cancer and prostate cancer posts. These include a wide variety of medical concepts. We can find treatments and medications like "Abraxane", a prescription medicine used to treat advanced cancer; generic cancer-related terminology like "chemotherapy", "tumor" or "metastasis";

medical procedures like "PSA", a screening test for prostate cancer, and "Gleason", a popular grading system used for prostate cancer prognosis; conditions like "incontinence", a common side-effect in treated prostate cancer patients; target body parts like "prostate" and "pancreas", as well as indirectly involved organs like "liver", which is strictly related to pancreatic adenocarcinomas, as more than 50% of patients with pancreatic cancer have liver metastases at the time of diagnosis [13].

Post Embeddings

The value of word embeddings in text classification is widely known nowadays [9]. Embeddings provide a dense vectorial representation of words and documents that considers both word order and context, preserving semantic rules and enabling more refined comparisons.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language representation model based on Transformers [15] designed to pre-train deep context-aware bidirectional representation of words. Since its development, this model has advanced the state of the art for many NLP tasks [4]. In its multilingual version, BERT enables the use of transfer learning for more than 100 languages, including Italian.

Word embeddings for each word of each post have been extracted averaging the last 4 layers of BERT, as described in [4]. For each post, word embeddings have been merged to obtain post embeddings, a document-level representation [10]. In this work we used the uncased multilingual $\text{BERT}_{\text{BASE}}$ configuration, which consists of 12 encoding layers, 12 attention heads and a hidden size of 768, adding a single linear layer at the end for sequence classification. Pre-trained embeddings have been fine-tuned on the validation set. We trained the model for 5 epochs, with early stopping based on F1-score, and we used Adam optimizer with a learning rate of 5e-5 and a batch size of 16. Post embeddings have been implemented in Python using PyTorch[3] and the Huggingface[4] Transformers library.

Text Classification

Each post has been binarily labeled for topic classification with no additional human supervision, leveraging the forum structure itself and the previous, implicit, annotation work done by moderators and users when they decide on which discussion subforum to post their messages. The representation methods described above have been combined in four ways for classification:

1. *$BERT^{FT}$*. No keywords, posts are represented using embeddings only. Pretrained representations have been fine-tuned (*FT*) on a validation set.
2. *Bag-of-Keywords*. Keywords match representation, posts are mapped using the BoK approach alone.
3. *$BERT^{FT}$+BoK*. Post embeddings are concatenated with BoK, resulting in a vector sized 768+*K* and made of both continuous and binary variables.
4. *$BERT^{PT}$+BoK*. Pre-trained (*PT*) post embeddings are concatenated with BoK without further fine-tuning.

Post representations in turn become the input of a logistic regression, regularized through Elastic Net to control the high number of variables. For *$BERT^{FT}$*, we directly used the output of the final layer instead. The whole pipeline, summarized in Figure 1, has been implemented[5] with Python 3 on a consumer machine's CPU.

[3] PyTorch website: https://pytorch.org/
[4] Huggingface website: https://huggingface.co
[5] Code available upon request

Results

Models have been evaluated in multiple configurations, varying the size of the keyword set *K*. For each configuration, the pipeline has been run 5 times with different seeds to monitor variability and stability of results. The truncation threshold *W* has been set to 380 tokens, obtained by using the WordPiece sub-word tokenization. This is equal to the 95th percentile of the MMB message length distribution after preprocessing and allows us to speed up computation by cutting only the tail of the distribution, shown in Figure 3.

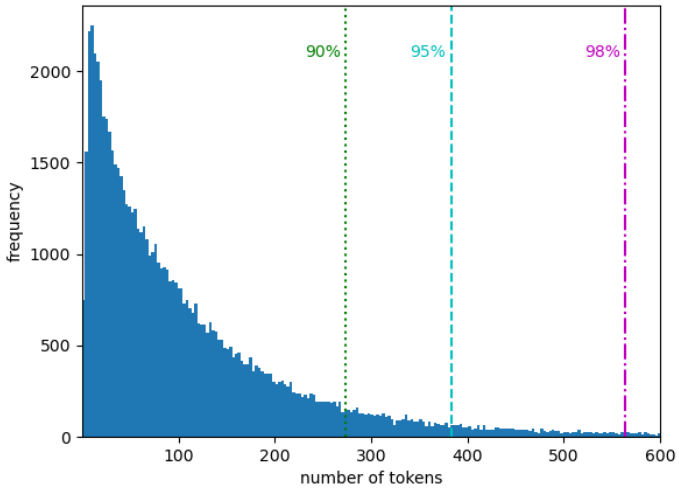

Figure 3 – Message length in terms of sub-word tokens

We evaluated models' performance with respect to their AUC-ROC, accuracy, F1-score and recall. Standard evaluation metrics have been complemented with recall to emphasize the prediction of actual positives. Experiments have been carried out on multiple topics to check the robustness of the models, changing the subforum of interest and the binary labeling accordingly.

Table 1 – Comparison of performance for the "Prostate Cancer" topic (W=380, M=5)

Model	K	AUC	Acc	Recall	F1
$BERT^{FT}$	0	.89±.03	.80±.03	.80±.05	.80±.04
$BERT^{FT}$+BoK	**5**	**.90±.02**	**.83±.03**	**.83±.02**	**.83±.03**
$BERT^{PT}$+BoK	5	.90±.02	.81±.03	.81±.03	.80±.03
BoK	5	.76±.03	.73±.05	.48±.08	.64±.06
$BERT^{FT}$+BoK	**50**	.90±.02	**.82±.03**	**.83±.03**	**.83±.03**
$BERT^{PT}$+BoK	50	**.91±.01**	.81±.01	.80±.02	.80±.02
BoK	50	.86±.03	.77±.03	.62±.03	.73±.03
$BERT^{FT}$+BoK	**200**	**.90±.02**	**.82±.03**	**.83±.03**	**.82±.03**
$BERT^{PT}$+BoK	200	.90±.01	.81±.01	.79±.02	.80±.01
BoK	200	.87±.01	.78±.02	.70±.05	.76±.02

Table 1 shows results for a short selection of *K* when *W*=380, *M*=5 and the target topic is "Prostate Cancer". The best-scoring configuration is *$BERT^{FT}$+BoK*. For this topic, both stand-alone *$BERT^{FT}$* and *$BERT^{PT}$+BoK* perform better than stand-alone *BoK*, whose performances grow with *K*.

Under the same configuration, we reported in Figure 4 the average F1-score trend for "Pancreatic Cancer" over all the tested values of *K*. Similar curves have been found for other topics of interest.

We also compared the amount of improvement between *BoK* and *BERT^{FT}+BoK* for different topics in terms of their average score and standard deviation, as shown in Figure 5. Our experiment highlights how *BERT^{FT}+BoK* is the best performing approach across the Aimac discussion board.

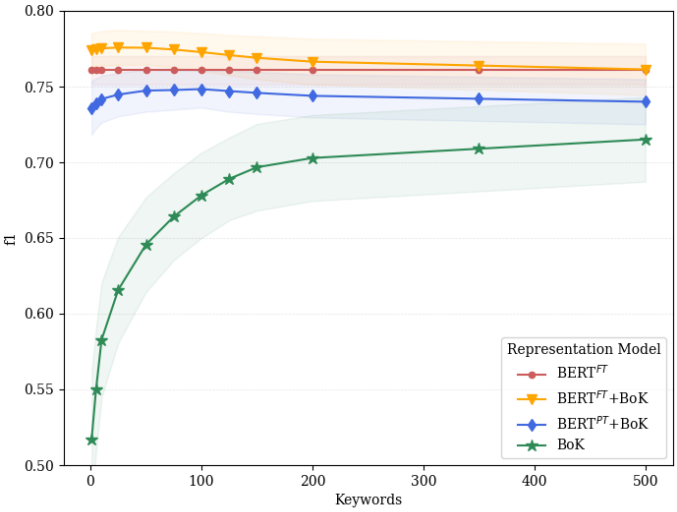

Figure 4 – F1 trend for "Pancreatic Cancer" (W=200, M=5). Shadows show the standard deviation for each model.

Discussion

Results show that introducing Multilingual BERT improves the quality of predictions when combined with keywords matching. Recall, in particular, is significantly higher in BERT-powered models than the BoK model itself. The magnitude of these improvements changes with the specific topic of interest. This may be influenced by the sample size of the target topic as well as the writing styles of the users, which may also affect the variability of the results.

While showing different performances, most of the topics share the same trend when varying *K*. The improvement of *BERT^{FT}+BoK* over *BoK* tends to shrink as the keywords set grows, suggesting that Transformer-based representations may suit situations where we deal with only a few keywords, like post filtering with short queries.

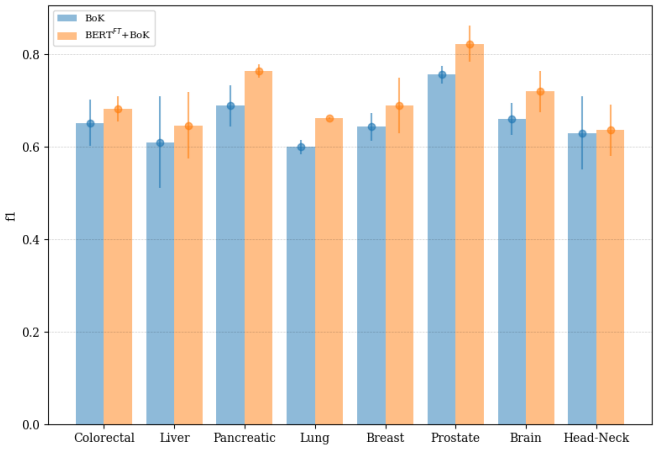

Figure 5 – Comparison of BERT^{FT}+BoK vs BoK in terms of F1-score (W=200, M=5, K=100). Whiskers highlight the standard deviation of the results for each topic.

It is interesting to observe the positive effect of fine-tuning in BERT models as in Figure 4, with *BERT^{FT}* and *BERT^{FT}+BoK* performing better than *BERT^{PT}+BoK*. The impact of fine-tuning depends on the topic and it is usually considerable, with some exceptions. In the "Lung Cancer" thread, for instance, fine-tuning does not seem to bring noticeable improvements over *BERT^{PT}*.

Comparing standalone *BERT^{FT}* and *BoK* models, we found that *BoK* is slightly superior to the Transformer-based representation in terms of accuracy in many topics, but widely inferior in terms of recall. This suggests that the first representation method, rigid by design, may be more suitable in scenarios where positive and negative detections are equally important, while the flexibility of embeddings does better at detecting positives by foregoing some points on the overall accuracy.

The combination of both BERT and BoK representations seems to overcome both the drawbacks, achieving a combination of high recall and accuracy. The improvements introduced by the *BERT^{FT}* component can be mainly attributed to two well-known capabilities of Transformer-based models: the ability to generalize, handling misspellings and ambiguities, and the attention mechanism, a powerful technique that directs the focus of the model only to the relevant parts of the input sequence. In this sense, *BERT^{FT}* uses attention to learn its own keywords, as exemplified in Figure 6. This figure shows a MMB Italian message about prostate cancer, highlighting the attention weights for each word. Attention weights are obtained by averaging the weights coming from the attention layers. The word with the highest attention is "prostate", followed by "psa" and "father". This is in line with expectations: the most important word is the involved organ itself, the second one is a prostate-related screening test, the latter implies demographic characteristics related to this type of cancer, that mainly affects men over 50.

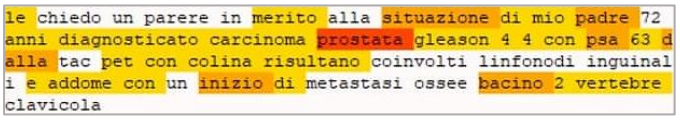

Figure 6 – Self-attention visualization of a MMB message about prostate cancer. The highlighted words are the ones with the highest weights, i.e., the self-learned keywords.

Limitations

Relying on the forum's own structure for automatic labeling is extremely convenient, avoiding manual reading and annotation of thousands of posts. However, this strategy implicitly assumes that every message in every thread under the subforum of interest indeed focuses on the target topic. This assumption is not always true: sometimes users join specific topics to express emotional support or gratitude, sometimes they share their experience about different topics. This introduces noise in labeling, which is partially addressed by keeping only the first *M* posts for each discussion.

In order to measure the impact of having noisy labels on the models' performance, we manually de-noised a few topics to check the extent of improvement over the same topics with noisy labels. Noise in labels has been removed by checking, for each message posted on the subforum board, whether the content was actually involving the subject of interest or not. Most of the time, de-noising leads to higher and stable performance, as shown in Figure 7 for prostate cancer. However, manual labeling is significantly time-consuming and cannot be considered a viable solution for large corpora.

Amongst useful terms like medications and oncological terms, the automatic keywords extraction through TF-IDF could introduce noise as well, giving importance to unrelated but recurrent terms like usernames and locations, as shown in Figure 2.

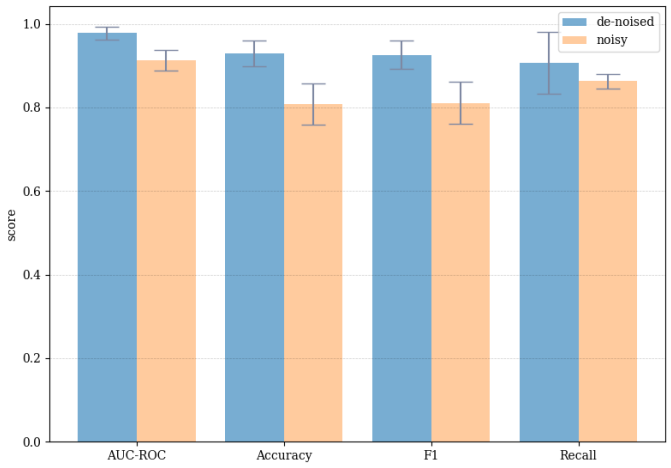

Figure 7 – Impact of noisy labels for the "Prostate Cancer" topic on AUC-ROC, accuracy, F1 score, and recall

We believe that the advantages of preserving an automated pipeline outweigh the drawbacks of introducing noise in the system for this experiment, where we maintain the focus on comparisons rather than absolute values.

Our investigations focus on multiple binary classifications implementing a One-vs-Rest strategy to inspect each topic separately. This could also be addressed as a multi-classification problem using different experimental setups.

Moreover, despite having more than thirty available subforums to perform the classification on, many of them are underrepresented and cannot be treated properly. Preliminary analyses on the "Nutrition" subforum, for instance, show how an insufficient sample size can affect both BERT and BoK stability, leading to excessively large fluctuations in results. This can already be noticed in Figure 5 for the "Head and Neck Cancer" or the "Liver Cancer" topics, where the final sample size (184 and 330 samples, respectively) is much smaller than that of "Lung Cancer" (2486 samples).

Conclusions

In this paper we presented a comparison between two different representation strategies for MMB topic classification. We demonstrated that combining simple strategies based on keywords matching with contextual embeddings like BERT seems to improve the quality of predictions for every topic we tried, providing a relatively inexpensive way to overcome BoW limitations and charting a course towards new feasible approaches for real-world situations involving languages other than English.

Acknowledgements

The work described in the article has been funded by the European Union's Horizon 2020 research programme under grant agreement No 875052 (CAPABLE, www.capable-project.eu).

References

[1] A. Benton, J.H. Holmes, S. Hill, A. Chung, and L. Ungar, medpie: an information extraction package for medical message board posts., *Bioinforma. Oxf. Engl.* 28 (2012) 743–4. doi:10.1093/bioinformatics/bts030.

[2] CAPABLE Consortium, CAPABLE website, (n.d.). https://capable-project.eu/ (accessed February 1, 2021).

[3] P. Cichosz, A Case Study in Text Mining of Discussion Forum Posts: Classification with Bag of Words and Global Vectors, *Int. J. Appl. Math. Comput. Sci.* 28 (2018) 787–801. doi:10.2478/amcs-2018-0060.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Vol. 1 Long Short Pap., Association for Computational Linguistics, Minneapolis, Minnesota, 2019: pp. 4171–4186. doi:10.18653/v1/N19-1423.

[5] S.T. Dumais, Latent semantic analysis, *Annu. Rev. Inf. Sci. Technol.* 38 (2004) 188–230. doi:https://doi.org/10.1002/aris.1440380105.

[6] A. Hulth, and B.B. Megyesi, A Study on Automatically Extracted Keywords in Text Categorization, in: Proc. 21st Int. Conf. Comput. Linguist. 44th Annu. Meet. Assoc. Comput. Linguist., Association for Computational Linguistics, Sydney, Australia, 2006: pp. 537–544. doi:10.3115/1220175.1220243.

[7] Italian Association for Cancer patients, relatives and friends, AIMaC website, (n.d.). https://www.aimac.it/ (accessed February 1, 2021).

[8] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, The State and Fate of Linguistic Diversity and Inclusion in the NLP World, in: Proc. 58th Annu. Meet. Assoc. Comput. Linguist., Association for Computational Linguistics, Online, 2020: pp. 6282–6293. doi:10.18653/v1/2020.acl-main.560.

[9] F.K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, A survey of word embeddings for clinical text, *J. Biomed. Inform. X*. 4 (2019) 100057. doi:10.1016/j.yjbinx.2019.100057.

[10] Q. Le, and T. Mikolov, Distributed representations of sentences and documents, in: Proc. 31st Int. Conf. Int. Conf. Mach. Learn. - Vol. 32, JMLR.org, Beijing, China, 2014: p. II-1188-II–1196.

[11] B.A. Lewin, and Y. Donner, Communication in Internet message boards, *Engl. Today*. 18 (2002) 29–37. doi:10.1017/S026607840200305X.

[12] J.J. Mao, A. Chung, A. Benton, S. Hill, L. Ungar, C.E. Leonard, S. Hennessy, and J.H. Holmes, Online discussion of drug side effects and discontinuation among breast cancer survivors, *Pharmacoepidemiol. Drug Saf.* 22 (2013) 256–262. doi:10.1002/pds.3365.

[13] H. Ouyang, P. Wang, Z. Meng, Z. Chen, E. Yu, H. Jin, D.Z. Chang, Z. Liao, L. Cohen, and L. Liu, Multimodality Treatment of Pancreatic Cancer with Liver Metastases using Chemotherapy, Radiation Therapy, and/or Chinese Herbal Medicine, *Pancreas*. 40 (2011) 120–125. doi:10.1097/MPA.0b013e3181e6e398.

[14] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, Understanding the Behaviors of BERT in Ranking, *ArXiv190407531 Cs*. (2019). http://arxiv.org/abs/1904.07531 (accessed February 3, 2021).

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in: Proc. 31st Int. Conf. Neural Inf. Process. Syst., Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 6000–6010.

[16] F. Wang, Z. Wang, Z. Li, and J.-R. Wen, Concept-based Short Text Classification and Ranking, in: 2014: pp. 1069–1078. doi:10.1145/2661829.2662067.

Address for correspondence

Tommaso Mario Buonocore, buonocore.tms@gmail.com.