GATTINA - GenerAtion of TiTles for Italian News Articles: A CALAMITA Challenge

Maria Francis[1,2,*,†], Matteo Rinaldi[3,†], Jacopo Gili[3,†], Leonardo De Cosmo[4], Sandro Iannaccone[5], Malvina Nissim[1,‡] and Viviana Patti[3,‡]

[1]CLCG, University of Groningen
[2]University of Trento
[3]University of Turin
[4]ANSA
[5]Galileo

Abstract

We introduce a new benchmark designed to evaluate the ability of Large Language Models (LLMs) to generate Italian-language headlines for science news articles. The benchmark is based on a large dataset of science news articles obtained from Ansa Scienza and Galileo, two important Italian media outlets. Effective headline generation requires more than summarizing article content; headlines must also be informative, engaging, and suitable for the topic and target audience, making automatic evaluation particularly challenging. To address this, we propose two novel transformer-based metrics to assess headline quality. We aim for this benchmark to support the evaluation of Italian LLMs and to foster the development of tools to assist in editorial workflows.

Keywords

CALAMITA Challenge, Italian, Benchmarking, Headline generation, Summarisation, LLMs

1. Introduction and Motivation

The title is undoubtedly one of the most important and crucial components of a journalistic article. A good title intrigues the reader, synthesises the news without anticipating its details, encourages further reading, and is simultaneously pleasant to read or hear. Often, the fate of an article is inextricably linked to the quality of its accompanying title: it is not uncommon for inherently interesting, in-depth, and factually correct articles to go unnoticed simply because they are accompanied by an inappropriate or unattractive title. Composing adequate titles is not a simple operation; it requires experience, sensitivity, balance, a sense of measure, and a deep understanding of the readers. There are no precise and inescapable "rules" – save, of course, for the usual deontological norms of pertinence and truth that regulate the journalistic profession – but in fact, the operation depends almost exclusively on the author's expertise and must be evaluated on a case-by-case basis.

Factors that can influence the composition of a title include, for example, the topic and the "tone of voice" of the article (a piece reporting a crime news story, for instance, requires a measured, discreet, and respectful title; conversely, a piece on lifestyle can and should be paired with a lighter, ironic, and more colorful title); the style of the publication hosting the article; the destination format (the same article printed in a paper newspaper and published on an online outlet, for example, typically has two different titles); potential "conflicts" with other titles present on the same page (for instance: repetitions of the same word or phrase, or the enunciation of contradictory concepts); space limitations; prescriptions related to search engine optimisation (for example, the use of a particular word or expression particularly popular at the time of publication, or a specific position of words within the title).

It is in this context that the journalist's toolkit has recently been enriched with a powerful new tool: Large language models (LLMs) undoubtedly have an important role in the world of journalism, including quality journalism. Although incapable of "understanding" content as a human journalist would, as well as the meaning of

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
† Shared first authorship.
‡ Shared supervision.
maria.francis@unitn.it (M. Francis); matteo.rinaldi@unito.it (M. Rinaldi); jacopo.gili584@edu.unito.it (J. Gili); leodecosmo@gmail.com (L. D. Cosmo); iannaccone@galileonet.it (S. Iannaccone); m.nissim@rug.nl (M. Nissim); viviana.patti@unito.it (V. Patti)
https://github.com/rosakun (M. Francis); https://github.com/mrinaldi97 (M. Rinaldi); https://github.com/Jj-source (J. Gili); https://github.com/malvinanissim (M. Nissim); https://github.com/vivpatti (V. Patti)
0009-0007-7638-9963 (M. Francis); 0009-0004-7488-8855 (M. Rinaldi); 0009-0007-1343-3760 (J. Gili); 0000-0001-5289-0971 (M. Nissim); 0000-0001-5991-370X (V. Patti)

words, LLMs are naturally capable of producing fluent, complex, plausible, and credible texts in a matter of moments. These models not only can improve the efficiency of editorial processes but also offer new creative and innovative possibilities for content creation, including the automatic generation of journalistic headlines. Analysing why it may be useful for journalism to have an LLM capable of generating titles leads us to consider numerous factors, such as time optimisation, content personalization, and the ability to maintain a high level of quality, coherence, and communicative impact. However, these tools also present many limitations and some dangers, particularly the risk of blindly relying on them.

Timing and speed, in particular, are one of the great challenges of journalism - being the first to publish a story, especially online, is often essential to attract readers - however, as we have seen, generating effective and incisive titles requires skill and time, which is not always available. An LLM can drastically reduce the time needed to create appropriate titles, for example by suggesting to the author a series of reasoned choices or proposing modifications and corrections to an already written title, always keeping in mind preset criteria such as length, tone, attractiveness, clarity, and the publication's style. Furthermore, if trained on the corpus of a particular publication, an LLM can suggest titles consistent with its tone of voice and editorial history.

Another important advantage that the use of LLMs can offer is the ability to personalise content for different platforms and audiences. In today's newsrooms, journalists no longer have to worry only about print media but must also consider the web, social media, newsletters, and other digital distribution platforms. Each platform requires a different type of language, style, and length for titles. For example, a title optimised for Twitter (or X) must be short and incisive, while a title for a news website can be more descriptive. An LLM is capable of generating variants of a title based on the medium of dissemination, allowing newsrooms to adapt their content precisely and in a targeted manner. Moreover, using reader behavioural data, the LLM can generate more attractive titles for specific demographic groups, thus improving the engagement and communicative effectiveness of the news.

With this task, which is developed in the context of the CALAMITA Challenge [1] and which consists in asking an LLM to generate a headline given the corresponding full article, we have a twofold aim.

The first aim is to test and analyse the ability of existing and future LLMs on the task of headline generation in the context of Italian news articles. This would provide a substantial step forward compared to past experiments on headline generation for Italian, which were run training much smaller sequence-to-sequence models from scratch [2, 3]. We expect that some of the shortcomings of the automatically generated headlines which were observed in previous work, such as lack of fluency and creativity [2], might not affect LLM-based generations.

The second aim is to provide a reliable, high quality dataset of articles and corresponding headlines in Italian, developed through a direct collaboration of language technology experts and journalists, which can be used and analysed well beyond the CALAMITA challenge. Although similar datasets exist for other languages [4, 5], this resource is still lacking for Italian.

Overall, experimenting with the use of LLMs for title generation can also be considered a first step towards the introduction of more extensive and comprehensive artificial intelligence agents, which assist the journalist in all phases of the creative process, from news research to drafting an outline, to writing the actual piece, and finally to its promotion. Indeed, a close interaction of language models and humans in this task has recently been shown to be key [6].

2. Challenge Description

The task of headline generation has often been treated as equal to an extreme summarization task [3, 7]. However, simply synthesising the content of the article into a brief description is not enough to provide a satisfying title. Additional characteristics such as attractiveness, creativeness, and many others also play a role. Writing appropriate headlines is challenging, even for current state-of-the-art LLMs.

Evaluating LLMs on the task of headline generation for Italian news articles thus serves multiple purposes. On one hand, it tests models' capacity to properly understand, that is, to reprocess large source texts in a way that is faithful to the content of the text. On the other hand, it acts as a means to assess the performance of LLMs in many complex dimensions, such as attractiveness, creativity, or adherence to tone. Finally, this benchmark could prove useful in practical applications. For instance, it may help guide decisions on whether, and to what extent, a journal should integrate LLMs into its workflow. It may also serve as an effective testbed for future research and development towards effective deployment in real-world scenarios - One such venue could be the use of prompting to achieve the desired style and tone in generated headlines.

In our challenge, language models are tasked with generating Italian-language headlines based on articles from scientific news journals written in Italian. Our dataset includes original articles from such journals, along with their human-authored titles. Models are provided the complete source text in the prompt, as well as instructions to generate a title that is brief, coherent, and captivating. We guide the model towards the specific editorial

style of the media outlet by including a small number of examples of headlines in our prompt. We employ automatic metrics that assess the model's performance along three dimensions:

1. Coherency with the original article (HA classifier)
2. Alignment with the style of human written headlines (NS classifier)
3. Similarity between the generated and the gold-standard headline (ROUGE [8], SBERT [9])

However, considering the complexity of the task, we believe that manually reviewing a sample of the generated headlines can offer additional perspectives on the behaviour of the model.

3. Data description

Our benchmark is based of two datasets consisting of science news articles from two different sources. In each dataset, we provide the full text of the article paired with the original, human-authored headline. Additionally, we include metadata such as link, date, author (if present) and subtitle.

3.1. Origin of data

The data were obtained via web scraping with custom Python scripts. Since links to articles more than a few weeks old are inaccessible on the Ansa website, we collected a large number by downloading the archived "Ansa Scienza" RSS feeds from The Wayback Machine and processing them to remove duplicates and extact links.

3.2. Data format

The data from web scraping were saved in "JSON Lines" (JSONL) format, with each line containing a JSON object with the following fields:

- **Title**: the title of the article
- **Source**: the name of the website
- **Date**: the publishing date of the article
- **Author**: the author of the article, if present
- **URL**: the Internet address of the article
- **Text**: the body of the article
- **ID**: a unique identifier of the article

3.3. Detailed data statistics

Our dataset consists of 30,461 articles gathered from two sources:

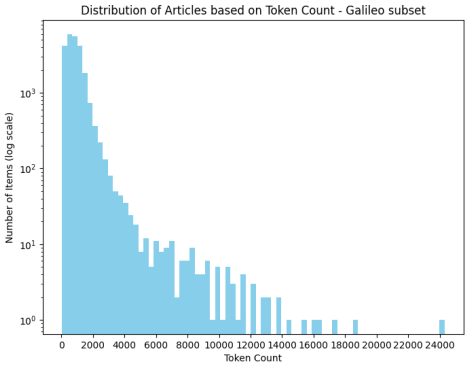

Figure 1: Distribution of articles by token count in the Galileo subset.

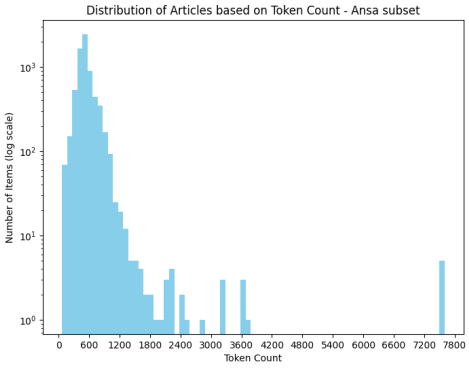

Figure 2: Distribution of articles by token count in the Ansa subset.

1. "ANSA scienza", the science section of the Italian newspaper "ANSA", from which obtained 6,889 articles: 649 of which are from 2024, and the others are from a period of time between 2018 and 2022.
2. The "Galileo" website, from which we sourced 23,572 articles dating from April 1996 to May 2024.

When measured with "tiktoken o200k_base" tokenizer model, we obtained a total of 21,365,897 tokens for the Galileo dataset (average: 906 tokens per article, maximum: 24,306) and a total of 3,762,539 tokens for the Galileo dataset (average: 546 tokens per article, maximum: 7,600). Figures 1 and 2 depict the distribution of articles by token count in the Galileo and Ansa datasets respectively.

3.4. Prompting

Due to the length of each article, the use of task examples in our prompt would be too computationally expensive. Therefore, we test the models in a zero-shot prompting setting. While we do not use any task examples in our prompt, we do provide seven examples of headlines. In this way, the model is given examples of the expected output (a title) rather than examples of the full task (article and title). Professional journalists made a list of 22 headlines that, in their opinion, were representative of a well-made writing process under the three aspects of being captivating, short and informative.

Each time the model is tested, 7 randomly chosen titles from the list are appended to the standard prompt. As a reference, the identifier of the example headlines is also saved along with the output of the model. See Box 1 for our input prompt.

Prompt for the LLM

Il tuo compito è generare un titolo accattivante e informativo per l'articolo fornito.
Requisiti:
- Titolo breve
- Cattura l'essenza dell'articolo
- Usa un linguaggio vivido e coinvolgente
- Non generare alcun tipo di testo che non sia il titolo dell'articolo
- Usa esclusivamente l'Italiano.
Presta particolare attenzione ai seguenti titoli di esempio e adotta lo stesso stile:
Title 1
Title 2
...
Title 7

Your task is to generate a catchy and informative title for the article provided.
Requirements:
- Short title
- Capture the essence of the article
- Use vivid and engaging language
- Do not generate any type of text other than the title of the article
- Use Italian exclusively.
Pay particular attention to the following example titles and adopt the same style:
Title 1
Title 2
...
Title 7

Box 1: Zero-shot prompt and English translation.

4. Preliminary Evaluation

To get a first impression of LLM performance on our task, we conducted preliminary experiments by manually reviewing headlines generated by several models. Overall, the results were unsatisfactory - while the titles were generally coherent with the articles, they lacked captivation and originality. The majority of the generated headlines followed the format *<Keywords: explanation>*, leading to repetitive and poorly formulated headlines. Examples of our preliminary results can be found in Table 1 in Appendix A. This behaviour persisted even when the models were explicitly instructed to avoid using colons in the titles, or when examples of titles were given. Out of 3,006 headlines generated by Phi-3.5 Mini-Instruct, 2,940 headlines contained a colon. We obtained similar results using Mistral-7B-Instruct-v0.3, Qwen2-7B-Instruct, gemma-2-9b-it and Italia-9B-Instruct-v0.1. Manual experimentation with the commercial LLMs Claude 3.5 Sonnet[1] and ChatGPT 4o[2] yielded the same behaviour:

- **Titolo originale:** Una rapina cosmica nell'ammasso di galassie dell'Idra
- **Claude:** Rapina cosmica: il furto di gas nell'ammasso dell'Idra
- **ChatGPT:** Rapina Cosmica: NGC 3312 Derubata di Gas nell'Ammasso di Galassie dell'Idra

Interestingly, when we asked Claude 3.5 Sonnet to improve our prompt for generating headlines, it added the line *<Struttura: [Frase d'impatto o dato interessante]: [Spiegazione o contesto]>* to our example prompt, explicitly requesting the unwanted behaviour. It appears that LLMs consistently regard this particular structure as the ideal format for a headline.

Given the inherent difficulty of interpreting LLM behaviour, we cannot provide a single reason for their preference for this particular construction. Of course, there might be a large presence of such headlines in the training data, particularly from lower-quality journals. There may also be an influence of Search Engine Optimizations (SEO) on the behaviour of the model: Giving importance to keywords is a classic SEO technique.

Moreover, we generally noticed a preference toward sentences poor in determinative and indefinite articles when compared with human written headlines.

5. Metrics

Automatically evaluating the quality of generated headlines is a challenging matter because headline quality is inherently subjective, multi-faceted, and context-dependent. Thus, instead of providing a single numeric

[1] https://www.anthropic.com/news/claude-3-5-sonnet
[2] https://openai.com/index/hello-gpt-4o/

value as an overall quality score, headlines should be evaluated along multiple dimensions and subsequently rated for their quality based on specific use cases. To give examples of what others have done - Cafagna et al. [2] evaluate generated headlines based on the criteria such as grammatical correctness, topic relevance, attractiveness, and overall appropriateness. Cai et al. [10] assess factors such as factual consistency, relevance, and surface overlap between the generated headline and the article, as well as its alignment with user-specific preferences.

In the aforementioned papers, the headlines were scored by human evaluators. This approach is resource intensive - to account for differences in individual preferences, hiring multiple human evaluators from varying demographic backgrounds is preferred. This does not scale well to the evaluation of multiple models on large-scale benchmarks across multiple studies, making the ability to automatically evaluate the outputs of LLMs essential.

Historically, n-gram overlap metrics like BLEU [11], ROUGE [8], or METEOR [12] have been used to compare generated outputs with reference "gold standard" texts, but these metrics emphasise surface-level matching and are therefore not robust to paraphrasing or other variations in acceptable outputs. Learned metrics such as COMET [13], a metric designed to mimic human quality judgement for machine translations, have been gaining in popularity. These are not easily transferable to other languages or tasks, and learnable metrics designed specifically for Italian headline generation are not available. Additionally, such metrics typically produce a single numerical score of 'quality'. To improve interpretability and ensure contextual flexibility, we would prefer to provide individual scores for each dimension. We train two novel learned metrics for Italian headline generation, but leave others for future work.

We evaluate model performance on our benchmark using four metrics: ROUGE [8], SBERT [9], and two custom metrics - the Headline-Article and Natural-Synthetic classifiers. Within the context of the CALAMITA challenge, the model's final score will be an aggregate in which four all metrics are weighted equally. Each metric is detailed in the following section.

5.1. ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [8] is a popular metric used to evaluate automatically generated summarizations. It provides a measure of overlap between generated text and gold-standard references. ROUGE is easily interpretable and allows for easy comparison across many papers due to its widespread use. However, it is not robust to variations in input, making it less suitable for the assessment of tasks involving creativity, such as headline generation. Following others [14], we will evaluate our system outputs using ROUGE-L, which identifies the length of the longest common subsequence between system and reference.

5.2. SBERT

Sentence-BERT, or SBERT [9], is a modification of the BERT network that uses Siamese networks and that can derive semantically meaningful, fixed-size vector embeddings from whole sentences. We use SBERT to compare our generated headlines to the gold-standard ones by comparing their SBERT embeddings using cosine similarity, which we then use directly as the similarity score. SBERT produces more meaningful sentence embeddings compared to BERT, which is not designed for sentence similarity tasks - therefore, cosine similarity with BERT embeddings could produce unwanted and less interpretable results.

5.3. Custom metrics

Given the limitations of the current available metrics for the headlines generation task, we develop two custom metrics employing classifiers based on Transformer [15] models. We trained both classifiers on a subset of the "blogs" section of the "Testimole"[3] dataset, which was obtained by web scraping various Italian media sources. Our subset consists of only those parts of the dataset scraped from professional media outlets. The criteria for the selection process, as well as the technical details for each classifier, are in Appendix B.

5.3.1. HA Classifier

Our first classifier is based on the Sentence Transformers [9] architecture, fine-tuned to discriminate between coherent and non-coherent pairs of headlines and articles. A generated headline can score between 0 and 1, representative of the degree of alignment between the headline and the content of the article. Following the work by De Mattei et al. [3], we call this classifier "HA", or Headline-Article.

To train the model, we used a non-finetuned Italian Sentence Bert model[4] to compute an embedding for each article. We then find the headline of the article in the dataset with the highest cosine similarity, and create a new dataset where each row contains the article (anchor), the original title (positive), and the title of the most similar article (negative). Because the original dataset contained some duplicate items, we filtered all articles with "1" as the cosine similarity score. With this dataset, we were able to use Triplet Loss to train the classifier

[3] https://huggingface.co/datasets/mrinaldi/TestiMole
[4] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

to differentiate between coherent and incoherent titles, starting from the assumption that the original title is the one most coherent with the article's content. We decided to perform a cosine similarity search instead of random shuffling in order to increase the difficulty of the discriminator's task.

The drawback of this approach is the low context window of the model - all articles were truncated after the first 512 tokens. While it is possible to develop a more complex architecture to account for larger texts, we leave this for future work.

5.3.2. NS Classifier

Our second classifier is called "NS", or Natural-Synthetic. It is a binary regression classifier based on an Italian BERT-base uncased model[5], trained to discriminate between human-authored and machine-generated titles. Given a title as input, the classifier outputs a numerical score indicating the likelihood of the title being close to those written by journalists. We believe that similarity to headlines written by journalists may be a useful indicator of the quality and appropriateness of a generated headline.

Using the same subset of Testimole employed for the "HA" classifier, we generated over 90,000 synthetic headlines using LLMs of up to 9 billion parameters. To avoid overfitting our classifier to the specific probability distribution of a single model, we generated synthetic headlines using different models; this process is detailed in Appendix C, along with details about the number of generated headlines per model. The result is a labelled dataset containing original as well as generated headlines.

The advantage of employing a "Natural-Synthetic" classifier is that the training objective is coarse, encouraging the classifier to consider a broad range of aspects that may account for the discrepancy of text generated by machines and humans.

6. Future works

We see value in future research using classifiers and regressors to assess specific aspects of generated headlines. Such metrics have the potential to capture complex probability distributions over a multitude of dimensions of the data, including dimensions that are not directly interpretable to human observation. For instance, a learned metric that predicts the amount of attention a headline will generated would be highly useful.

Inspired by Generative Adversarial Networks (GANs), we find the employment of classification-based metrics promising for developing a model specialized in headline generation. A discriminator/generator training system allows us to build a positive feedback loop in which the headline generation system teaches itself to generate good headlines based on the classification of the discriminator. For instance, the model can be trained to 'fool' the NS discriminator as often as possible while the NS discriminator uses the experience to improve at identifying synthetic data, causing both models to improve simultaneously. This method, for instance, should quickly solve the frequent use of the colon in automatically generated headlines outlined in Section 4.

7. Limitations

Our benchmark is limited to articles and headlines from only two journals, which restricts its representativeness across journalistic domains. As a result, it may not capture the variability present in publications targeting different demographics, covering varied topics, or representing a full spectrum of political perspectives.

In training our classifiers, we took care to prevent data contamination by ensuring non-overlapping splits between training and test sets. Nonetheless, given the public availability of the articles online, there remains a possibility that some test data may indirectly overlap with training data due to external access and prior exposure.

8. Ethical issues

This task is aimed at testing the factual knowledge which LLMs acquire during their training process, whose objective is language modelling. This task should not suggest, or stimulate, that LLMs should commonly be used as knowledge bases or as reliable sources of factual information. The investigation underlying this challenge is research-oriented, aimed at a better understanding of LLMs' abilities, and possibly suggest ways to discern when models might be providing more or less reliable knowledge and possibly making them more transparent in their generated output.

9. Data license and copyright issues

Access to the data is granted for the evaluation but cannot be shared publicly at the moment, also for reasons related to data contamination.

Acknowledgments

The authors would like to thank *ANSA Scienza* and *Galileo, giornale di scienza* - http:\www.galileonet.it for

[5] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

their interest in the GATTINA CALAMITA challenge and for the extremely valuable exchange of ideas that allowed us to shape a task of high potential impact in the field of journalism.

References

[1] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[2] M. Cafagna, L. D. Mattei, D. Bacciu, M. Nissim, Suitable doesn't mean attractive. human-based evaluation of automatically generated headlines, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2481/paper13.pdf.

[3] L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, Invisible to people but not to machines: Evaluation of style-aware headlinegeneration in absence of reliable human judgment, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 6709–6717.

[4] X. Ao, X. Wang, L. Luo, Y. Qiao, Q. He, X. Xie, Pens: A dataset and generic framework for personalized news headline generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 82–92.

[5] Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, et al., Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation, arXiv preprint arXiv:2004.01401 (2020).

[6] Z. Ding, A. Smith-Renner, W. Zhang, J. Tetreault, A. Jaimes, Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 3321–3339. URL: https://aclanthology.org/2023.findings-emnlp.217. doi:10.18653/v1/2023.findings-emnlp.217.

[7] A. Rush, A neural attention model for abstractive sentence summarization, arXiv Preprint, CoRR, abs/1509.00685 (2015).

[8] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[9] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[10] P. Cai, K. Song, S. Cho, H. Wang, X. Wang, H. Yu, F. Liu, D. Yu, Generating user-engaging news headlines, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 3265–3280.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[12] A. Lavie, M. J. Denkowski, The meteor metric for automatic evaluation of machine translation, Machine translation 23 (2009) 105–115.

[13] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, Comet: A neural framework for mt evaluation, arXiv preprint arXiv:2009.09025 (2020).

[14] M. Krubiński, P. Pecina, Towards unified uni-and multi-modal news headline generation, in: Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 437–450.

[15] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[16] M. Rinaldi, Testimole, 2024. URL: https://huggingface.co/datasets/mrinaldi/TestiMole.

A. Examples of Good titles selected by professional journalists

- Nella Via Lattea c'è un oggetto misterioso, è velocissimo
- Nasce il gemello digitale del rischio ambientale in Italia
- I cinque modi in cui il cervello invecchia
- Covid-19, il mistero degli over 90
- A 44 e a 60 anni i due gradini chiave dell'invecchiamento
- Palestra o snack? la scelta dipende da un messaggero chimico
- Dagli stadi alle spiagge, sono i salti a sincronizzare il ballo
- Dalle rose alle melanzane, ecco i geni delle spine
- Così il Covid accelera l'invecchiamento
- Uno zucchero naturale contro la calvizie, bene i test sui topi
- Scoperto nel cervello il circuito dell'effetto placebo
- Pronto il Google Earth del cuore umano
- Una molecola può ringiovanire il sistema immunitario
- Scoperto il dizionario dei sinonimi e contrari del cervello
- Le farfalle nello stomaco non sono solo un modo di dire
- Pronto il primo orologio nucleare, il più preciso del mondo
- Gli uccelli in volo si comportano come gli atomi
- L'Italia ritenta la sfida impossibile della geometria
- Le auto nel traffico come i batteri in cerca di cibo
- Robot come alleati, trovata la chiave per collaborare con gli umani
- Dalle spugne di vetro grattacieli più sottili e resistenti
- L'IA non è razionale, fa ragionamenti non logici

B. Examples of Model Outputs in Preliminary Evaluation

Original Title (Italian)	**LLM Generated Title (Italian)**
No allo studente flessibile *No to the flexible student*	Riforma Universitaria: Studenti in Rivolta contro la Selettività e la Flessibilità *University reform: Students in revolt against selectivity and flexibility*
Alzheimer, la proteina della speranza *Alzheimer, the protein of hope*	Nicastrina: nuova chiave per Alzheimer, scoperta in Calabria *Nicastrin: new key for Alzheimer's, discovered in Calabria*
Un laser estremo *An extreme laser*	Euv: Nuova Frontiera nella Manipolazione Molecolare con Luce Ultra-Sottile *EUV: New frontier in molecular manipulation with ultra-thin light*
E' italiano il primo modello genetico *The first genetic model is Italian*	Creazione topi per studiare il ritardo mentale: scoperta genetica rivoluzionaria *Creation of mice to study mental disability: revolutionary genetic discovery*
Carta d'identità per le balene *Identity card for whales*	Genetica della balena: Identificazione di specie protette nel commercio illegale *Whale genetics: Identification of protected species in illegal trade*
Scampati allo shock anafilattico *Escaped from anaphylactic shock*	Scoperto il fattore eNOS nella reazione anafilattica: speranza per una cura *eNOS factor discovered in anaphylactic reaction: hope for a cure*
La casa in un clic *The house in a click*	Domotica: Software unificato per controllare tutti gli elettrodomestici, innovazione Cnr *Home automation: Unified software to control all appliances, Cnr innovation*
Deforestazione senza fine *Endless deforestation*	Deforestazione in aumento in Amazzonia: il Brasile reagisce *Deforestation increasing in the Amazon: Brazil reacts*
"Più impegno contro il fumo" *More commitment against smoking*	Bioetica del tabagismo: raccomandazioni per campagne di prevenzione e interventi regolativi *Bioethics of smoking: recommendations for prevention campaigns and regulatory interventions*
Una carovana contro l'Oleoducto *A caravan against the pipeline*	Protesta contro l'Oleodotto Crudos Pasos: Ecuador, ambientalismo e scontro politico *Protest against the Crudos Pasos pipeline: Ecuador, environmentalism and political clash*
Analisi dell'egemonia americana *Analysis of American hegemony*	Egemonia Americana: Stati Fuorilegge e Diritto Internazionale *American hegemony: Rogue states and international law*
Censura sui dati scientifici *Censorship on scientific data*	Scienza e Sicurezza: Accademia e Governo Discutono su Divulgazione nei Lavori Scientifici *Science and security: Academia and government discuss disclosure in scientific works*
Siamo tutti africani: ricostruito il Dna degli antenati *We are all Africans: reconstructed the DNA of ancestors*	DNA dell'Africa: Prove definitive dell'origine umana *DNA of Africa: Definitive proof of human origin*
Rubbia racconta Galileo divulgatore di scienza *Rubbia tells Galileo as a science communicator*	Galileo Divulgatore: Scienza e Civiltà dal XVII al Mondo Moderno *Galileo communicator: Science and civilization from the 17th century to the modern world*
Galileo Galilei nel pensiero di Thomas Khun *Galileo Galilei in the thought of Thomas Khun*	Kuhn e Galileo: Rivoluzione scientifica attraverso la meccanica e l'esperimento *Kuhn and Galileo: Scientific revolution through mechanics and experiment*
Nella maglie della lingua *In the mesh of the language*	Realtà e Reti: L'Illusione di una Lingua Universale nella Scienza Cognitiva *Reality and networks: The illusion of a universal language in cognitive science*
Europa-America, la migrazione dei virus *Europe-America, the migration of viruses*	Scambi e mutua scoperta: Storia sanitaria dell'unificazione globale *Exchanges and mutual discovery: Health history of global unification*
Infiniti universi allacciati *Infinite universes connected*	Inflazione cosmica: un universo di mondi nascosti *Cosmic inflation: a universe of hidden worlds*

Table 1
Comparison of Original and LLM Generated Titles with Literal Translations.

C. Composition of the datasets used to train the classifiers

The dataset we used as a source of material for both the NS and HA classifiers is taken from "Testimole" [16], a massive collection of Italian web scraping data that includes a "blogs" subset containing, as of November 2024, more than 2.8 million posts from various online blogs and websites. From the original 2.8 million rows, we obtained a much smaller dataset by filtering articles coming from sources that are, to our judgement, more similar to professional media outlets. After this selection process, which yielded a total of 715,335 articles, we filtered out articles written in languages different than Italian by using the "FastText Lang ID" field already present in Testimole. After the foreign-languages pruning the count of articles was 293,518 articles. Finally, we discarded all the rows whose article was shorter than 350 characters to arrive to a final dataset size of **264,455 articles**. **In the following section, this dataset will be referred as "*testimole-subset*".** In order to increase the diversity of data for the HA Classifier, we added to this dataset a collection of 432.000 articles taken from the professional Italian media outlet "Il Fatto Quotidiano": we had to add this source manually because the articles were missing from the original Testimole dataset due to a scraping issue. In the section of HA Classifier, we will refer to this additional subset as "*testimole-subset-auxiliary*". Finally, we are going to refer to the small subset of Galileo used in the testing process as "*experimental-dataset*". The experimental dataset contains 3007 original headlines from "Galileo" and 3007 headlines generated using Phi 3.5 Mini Instruct from the same subset of Galileo's articles.

D. NS Classifier

For the NS Classifier, we decided to split the *testimole-subset* dataset in two sets: 60% of the dataset was kept with the original headline ("*natural*") while in the remaining 40% the original headline was substituted with a generated one ("*synthetic*"). The original headline is kept as a reference as a separate column in the dataset. Specifically, we generated 93,921 headlines and kept 132,227 original headlines. There is no contamination between generated and original headlines: no synthetic headlines were generated for headlines that are present in the dataset with the "natural" label. The dataset was then divided in "test" (45230 entries, x natural, x syntethic) and "train" (180918 entries, 105885 natural, 75033 synthetic) split for training. For the generation, we ran Ollama on different models using the same prompt adopted for the evaluation. In Table 2 you can see the amount of generated headlines for each model used.

The classifier was created using Hugging Face's `transformers` library. We initialized the model using `AutoModelForSequenceClassification` and trained the model using a binary cross-entropy loss function (`BCEWithLogitsLoss`).

Training was conducted with a batch size of 32, a learning rate of 2 × 10^5}, and a warmup ratio of 0.1 to help stabilize early training. A linear learning rate scheduler and the `$AdamW$` optimizer with gradient clipping were employed to manage learning stability. We also implemented early stopping, monitoring the F1 score to save the best model checkpoint and halt training if the model failed to improve over multiple epochs. The resulting model obtained a 95% of accuracy on the test set. Accuracy is measured as the number of correctly guessed labels divided for the total number of examples. The threshold to decide for a positive or negative label was set at 0.5. Using a continuos score instead of the threshold led to the same result, for this reason we decided to kept only accuracy in this report.

After having tested the model, we decided to further train it on the test set in order to have an improved model to be used for the CALAMITA task.

We then tested this further trained model on the smaller "experimental-dataset" dataset containing 3007 natural and 3007 synthetic headlines coming from the Galileo dataset. This evaluation obtained an accuracy of 87%

While initially we directly used PyTorch to train the experimental versions of the model, we then decided for simplicity to adopt the HuggingFace transformer library to easily upload the model on the HuggingFace hub. The further trained version of model is available at the address: https://huggingface.co/mrinaldi/flash-it-ns-classifier-fpt

E. HA Classifier

In order to build the HA Classifier we first computed, for each article contained in the "testimole-subset" dataset, the embedding of the article's text using SentenceBert with an Italian model [6] and added the embedding to a new column in the dataset. Then, we paired each article (source) of the dataset with the article (target) having the highest cosine similarity between the embeddings. After the pairing, both source and target were marked as "used" so that each article can appear no more than one time in the resulting dataset, either as a source or as a target. The resulting dataset [7] has 6 columns:

- **Anchor**: the body of the "source" article
- **Positive**: the original title of the "source" article

[6] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

[7] https://huggingface.co/datasets/mrinaldi/flash-it-ha-dataset-cossim

Model	Count	Percentage
lama3.2:3b-instruct-fp16	51886	55.24%
qwen2.5:7b-instruct-q8_0	18418	19.61%
aya:8b-23-q8_0	17043	18.15%
mistral:7b-instruct-v0.3-q6_K	6312	6.72%
phi3.5:3.8b-mini-instruct-fp16	262	0.28%

Table 2
Distribution of generated headlines by model

- **Negative**: the original title of the "target" article
- **Cosine similarity**: the Cosine Similarity between the source's and target's embeddings computed on their texts
- **Url positive**: the URL of the source article, it can be used as a key to find the original article in the Testimole dataset
- **Url negative**: the URL of the target article

Given the procedure employed for generating this dataset, the resulting number of row is halved so that, starting from the original 256530 entries in the "testimole-subset" dataset we obtained 128265 entries, divided into 102600 train entries and 25665 test entries. We believe that using the cosine similarity instead of randomly shuffling the articles can improve the performance of the classifier by increasing the difficulty of the task. Results with a classifier trained on randomly paired articles is present in the table below.

The classifier was created using Sentence-BERT, specifically by initializing the model with the `SentenceTransformer` class from the `sentence_transformers` library, using a pre-trained Italian model[8]. To fine-tune this model, we employed a TripletLoss function to enhance similarity-based ranking in embedding space. The triplet loss was the optimal choice given our dataset because it requires an anchor, a positive and a negative example. The goal of the triplet loss is to maximize the distance between the anchor and the negative example while at the same time minimize the distance between the anchor and the positive example. In this way, we encouraged the formation of meaningful embeddings tailored to minimize the distance between an article and a title coherent with its content, notwithstanding the 512 token length limitation.

Training was conducted over three epochs with a batch size of 64 for training and 16 for evaluation, using a learning rate of 2×10^{-5} and a warmup ratio of 0.1 to stabilize initial training steps. We used the `$SentenceTransformerTrainingArguments$` to configure training, applying half-precision floating-point (fp16) to speed up processing. An evaluation was performed every 1,000 steps to monitor model performance, with checkpoints saved periodically to retain the best-performing model. We kept the "margin" value at "5" following the documentation of SentenceBert. [9]

The resulting classifier outputs a score representing the alignment between the article and its headline.

After having trained the HA Classifier on the "testimole-subset" dataset, we decided to use an additional dataset (testimole-auxilliary) to further improve the classifier. Testimole-Auxiliary, halved due to matching, has 216562 articles of which 108281 were used as train and 108281 as test. The same procedure used for *testimole-subset* was applied to *testimole-auxilliary*. In the following page we present a table summing up the results of the various models on the test datasets.

[8] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

[9] https://sbert.net/docs/package_reference/sentence_transformer/losses.html#tripletloss

Model name	Model training set	Test set	Correct Triplets	Accuracy	Avg pos. dist.	Avg neg. dist.	Average Margin	ROC AUC
HA-Cossim	"testimole-subset" (Train)	"testimole-subset" (Test)	21949	0.8552	0.4	0.73	0.33	0.84
HA-Cossim-FPT	"testimole-subset" (Train+Test)	"testimole-auxiliary" (Test)	98913	0.9135	0.37	0.72	0.35	0.89
HA-Cossim-FFPT	"testimole-subset" (Train+Test), "testimole-auxiliary" (Train)	"testimole-auxiliary" (Test)	106662	0.9850	0.3	0.76	0.47	0.96
HA-RANDOM	"testimole-subset" (Train)	"testimole-auxiliary" (Test)	92523	0.8545	0.24	0.40	0.16	0.8

Table 3
Report of the results obtained by HA Classifier on the test datasets