SciTechBaitRO: ClickBait Detection for Romanian Science and Technology News

Raluca-Andreea Gînga, Ana-Sabina Uban
gingaraluca@gmail.com, auban@fmi.unibuc.ro
Faculty of Mathematics and Computer Science,
Human Language Technologies Research Centre,
University of Bucharest

Abstract

In this paper, we introduce a new annotated corpus of clickbait news in a low-resource language - Romanian, and a rarely covered domain - science and technology news: SciTechBaitRO. It is one of the first and the largest corpus (almost 11,000 examples) of annotated clickbait texts for the Romanian language and the first one to focus on the sci-tech domain, to our knowledge. We evaluate the possibility of automatically detecting clickbait through a series of data analysis and machine learning experiments with varied features and models, including a range of linguistic features, classical machine learning (ML) models, deep learning and pre-trained models. We compare the performance of models using different kinds of features, and show that the best results are given by the BERT models, with results of up to 89% F1 score. We additionally evaluate the models in a cross-domain setting for news belonging to other categories (i.e. politics, sports, entertainment) and demonstrate their capacity to generalize by detecting clickbait news outside of domain with high F1-scores.

1 Introduction and Related Work

Clickbait is a form of content used with the intention of attracting as many readers as possible through a type of content supported by a specific title, designed to attract as many clicks as possible.

News media is no stranger to this way of attracting readers. Furthermore, the technique has been used for more than 100 years under the name of "yellow journalism" or "yellow press" (Britannica, 2024), i.e. that type of journalism that used shocking, "sensational" headlines to attract readers to buy the newspaper, without the news being interesting or at least partially supported by real facts.

Clickbait is used nowadays by news publications to promote articles on social networks (i.e. Facebook, Instagram) by engineering news titles to contain certain terms, words, and patterns that arouse curiosity or revolt, such as "Incredible", "You must read this" or "It is outrageous" etc. This is to the detriment of the readers, who are being manipulated into clicking misleading links. Given how widespread this practice is and the amount of news published daily, automatic solutions for detecting clickbait can be a welcome solution. Some social media channels, such as Facebook, have already implemented a protocol to reduce clickbait content.

Technology and science play a crucial role in shaping modern society, driving progress, and improving the quality of life. From medical advancements that extend and save lives to innovations in communication that connect people across the globe, the impact of science and technology is profound and far-reaching. Socially, these fields are essential for addressing some of the most pressing challenges of our time, such as climate change, health crises, and sustainable development. By providing the tools and knowledge needed to understand and solve complex problems, science and technology empower societies to make informed decisions, promote economic growth, and enhance social equity.

In essence, the advancement of technology and science not only drives progress but also ensures that society can adapt, thrive, and respond effectively to the evolving needs of its members. In the technology and science domain, disinformation can be especially harmful by twisting scientific results and the public's trust in science. Recent examples such as research on the COVID-19 virus and vaccines have shown that misinformation about scientific findings can easily spread through manipulation methods such as clickbait and can be profoundly harmful for society.

One of the first studies on clickbait detection using machine learning techniques was published by Potthast et al. (2016), where standard ML models were used, including logistic Regression (LR), Naive Bayes (NB), Random Forest (RF). The study

Proceedings of the Third Workshop on NLP for Positive Impact, pages 188–201

was conducted on a compilation of a corpus of 2,992 English tweets, among which 767 were in the clickbait category. A novel contribution brought by this study is the generation of 215 independent variables that were further split into three types: teaser message aimed at capturing the characteristics of the clickbait teaser message (bag-of-words features, sentiment polarity, tweet's readability, use of Terrier stopword list (Ounis et al.), list of the easy words Dale and Chall, use of contractions, punctuation use); link-based features (analysis of web pages concerning tweets); meta information (encoding the tweet's sender, attaching an image/video to the tweet, retweet, the part of the day the tweet was sent).

Chakraborty et al. (2016) investigated several lexical and syntactic features for clickbait, achieving remarkable performance. A set of 15,000 article titles were analyzed with various features created and divided into various categories: sentence structure (average word length, title length, etc.), word patterns (existence of special punctuation patterns), clickbait language (standard and classical phrases from clickbait domain, slang, jargon (Ghanem et al., 2020)), N-gram features. More recent studies use deep learning (Gamage et al., 2021; Jain et al., 2021) for classifying clickbait for the English language.

Few studies on clickbait detection were performed for languages other than English. One notable example includes studies on clickbait detection performed for Turkish (Geckil et al., 2018) by forming and expanding a Turkish language dataset - ClickbaitTR (Genç and Surer, 2021), as well as Indonesian language, with the release of CLICK-ID (William and Sari, 2020).

For the Romanian language, on the other hand, the clickbait sphere has not been studied extensively. The only reserach in this direction is the very recently published study of Broscoteanu and Ionescu (2023) introducing RoCliCo, a general domain corpus specifically designed for clickbait in Romanian. Păcurar and Oprişa (2023) perform experiments on a previous version of our dataset, obtaining a 0.85 F1-score with a multi-layer perception classifier.

Our contributions in this paper include the release of an annotated corpus for the Romanian language on clickbait in the science and technology area, SciTechBaitRO - the first corpus for clickbait detection for science and technology news, and the largest clickbait corpus for Romanian, including approximately 11,000 samples. We apply various artificial intelligence algorithms in order to automatically predict clickbait titles, and show that detection is possible with an F1-score of 90%. Section 2 describes the methodology used for building our corpus, starting from the details related to the dataset collection and annotation, to data analysis and duplicate detection. In section 3, we describe a series of machine learning experiments performed to automatically detect clickbait based on the constructed corpus. We use a variety of models from classical ML models to deep learning and pre-trained transformers. We create linguistic-based features and measure the performance of models using these features in comparison to the pre-trained models using simple word sequence features. Section 4 reports the results obtained from the models, followed by experiments to measure the performance of the best model on out-of-domain news data (news belonging to other domains such as politics, economics) in Section 5. Finally, the last section (Section 6) concludes the study and offers some perspectives on what could be studied further in this direction.

2 Corpus Construction

Given the lack of annotated datasets in the field of news articles in the scientific-technology category, we release an annotated corpus of clickbait news in these domains for the Romanian language. It represents the first dataset of this kind, including annotated clickbait news for the Romanian language, SciTechBaitRO[1].

We collected a number of 10,867 articles from the scientific-technology area published on Romanian news websites between 1.02.2021 and 1.02.2022 and manually annotated the type of article (clickbait or not). We obtained 5,464 titles identified as non-clickbait and the remaining 5,403 in the clickbait category.

The following subsections discuss in more detail the methodologies used for collecting and annotating the data.

2.1 Data Collection

For the niche of science and technology, four main content publications and well-known news websites from Romania were chosen: Digi24[2], Play

[1] https://www.kaggle.com/datasets/andreeaginga/clickbait

[2] https://www.digi24.ro/stiri/sci-tech

Tech[3], Go4IT[4], Descopera.ro[5].

These websites are among the most popular news portals in Romania, presenting the main advantage of offering science and technology news on the latest innovations in IT&C field, scientific discoveries, news from the world of technology, gadgets, travel, and general culture.

In order to collect this data, we used web scrapers to collect news published on these websites between 1^{st} of February 2021 and 1^{st} of February 2022. The websites allow filtering of the news based on a category or tag, which we used to filter only science and technology news. Finally, we extracted the titles as well as the article body (full text of the news) and any associated keywords (tags that are part of some news for better Search Engine Optimization), and dates of publication, for all articles published in these categories in the specified timeframe. We record all this information in the published corpus. Table 1 shows examples of clickbait and non-clickbait language.

2.2 Data Annotation

The annotation was done manually by the authors of the paper. The annotators are Romanian native speakers, graduates of Romanian universities, with educational backgrounds ranging from Master's degrees to PhDs, ensuring a deep understanding of both the language and the nuances required for accurate labeling.

While some of the titles are very straightforward to identity manually as clickbait based on simple criteria (such as the first in the list below), we find a significant minority of corner cases which do not easily fit any specific pattern and can be ambiguous with regards to their correct label. We use the following main criteria for deciding if a title is clickbait, starting with the simple patterns and continuing with more delicate criteria used in the case of ambiguous examples:

Some of the criteria considered for annotation with the label 1 (clickbait) were:

- if the questions "When", "Who", "Where", "How" appeared in the title and are not answered,
- if there are terms, words that are meant to dramatize, to highlight the sensational, the incredible,
- if there are questions in the title asked with the purpose of making a reader curious and have the instinct to click through the article, which were not answered in the title (in case the title refers to more consistent information which could not fit in the title alone, we do not consider this a malicious omission that makes it clickbait)
- inspection of the article's content in comparison with the title showing any misleading statements in the title

In order to validate the annotations, we used ChatGPT (based on OpenAI's GPT3.5[6]) as a second annotator and identified instances where it disagreed with the first human annotator. The prompt used can be found in the Appendix (Section A.1) - we instructed ChatGPT to use the same criteria as was used by the human annotator. The obtained Cohen's kappa coefficient for the agreement between the human annotator and ChatGPT was 0.316. Finally, for the disagreement cases, a second human annotator independently annotated the examples and the final label was decided by the majority vote. In 69% of cases of initial disagreements, the second human annotator agreed with the first human annotator rather than with ChatGPT. The final agreement between the two human annotators on the subset of 3,781 annotated by two humans was 0.365.

2.3 Duplicate Detection

As a final data cleaning step, we eliminate duplicated news from the dataset. Since the news is sourced from the same period across several different websites, it is possible that the same news is posted multiple times or copied across websites, which can be a source of noise for our task.

We first identify exact duplicates and remove them from the dataset. We find 227 exact duplicates taking into account the entire body of the news article. When we also exclude duplicated news in terms of headlines from our dataset, we are left with a total of 10,640 news.

As a final precaution, we investigate whether, aside from the identical duplicates, there are any nearly identical duplicates that should be removed from the dataset. In order to identify highly similar news pairs, we use a SentenceBERT (Reimers and Gurevych, 2019) model to embed all news contents

[3] https://playtech.ro/tehnologie/
[4] https://www.go4it.ro/content/
[5] https://www.descopera.ro
[6] https://platform.openai.com/docs/models/gpt-3-5-turbo

Type	Headline	Translation
Clickbait	Cum arată cea mai rapidă ambulanţă din lume	What does the fastest ambulance in the world look like?
Non-clickbait	Oamenii de ştiinţă au demonstrat că materia poate deveni invizibilă	Scientists have proven that matter can become invisible

Table 1: Examples of titles classified as clickbait and non-clickbait

in our dataset. Specifically, we use Multilingual Sentence BERT which includes support for Romanian (Reimers and Gurevych, 2020). We compute similarities between pairs of headlines using cosine similarity on their respective embedding representations. We set a high threshold of 0.9 as to identify suspicious news pairs. This is followed by a manual evaluation step in which we verify whether the suspicious pairs appear to be duplicates or cases of plagiarism. We observe, in many cases, the articles selected as suspiciously similar were posted in different periods of time, or have particular differences that make them unique, despite the high similarity score. We conclude that none of the near duplicates selected based on sentence similarity scores seem to be duplicated or plagiarized, so we don't remove any articles at this step.

2.4 Exploratory Data Analysis

In Table 2, we see the distribution of clickbait and non-clickbait headlines. Of the 10,640 total headlines, 49.55% are classified as clickbait, while 50.45% are non-clickbait. This indicates that non-clickbait headlines are slightly more common in the dataset and that we have a balanced dataset.

Label	Count
Clickbait	5272
Non-clickbait	5368

Table 2: Classes distribution

Figure 1 highlights the POS distribution across clickbait and non-clickbait headlines. We observe notable differences in the usage of certain parts of speech between the two categories:

- The use of interjections (INTJ) is strikingly high in clickbait titles at 80.77%. This reflects a reliance on emotional or attention-grabbing expressions designed to elicit strong responses from the audience, often seen in exclamatory phrases like "Incredible!" or "Unbelievable!"
- Pronouns (PRON) also appear much more frequently in clickbait headlines (71.41%), indicating a strategy to create a personal connection with the reader. This use of pronouns, such as "you" or "your," helps to foster an intimate and engaging tone.
- Adverbs (ADV) are significantly more prevalent in clickbait headlines, with 62.42% of clickbait titles employing them. This suggests that clickbait often uses adverbs to emphasize emotional or sensational aspects of the content, enticing readers to engage more deeply.
- Verbs (VERB) are found more frequently in clickbait titles (52.33%) compared to non-clickbait titles (47.67%). This indicates a focus on action-oriented language that encourages immediate engagement, often prompting readers with phrases like "Find out how..." or "Discover the truth...".
- Conversely, numbers (NUM) are more prevalent in non-clickbait titles (59.77%), suggesting that these headlines are more focused on providing factual, data-driven information, appealing to readers looking for substantive content rather than sensationalism.

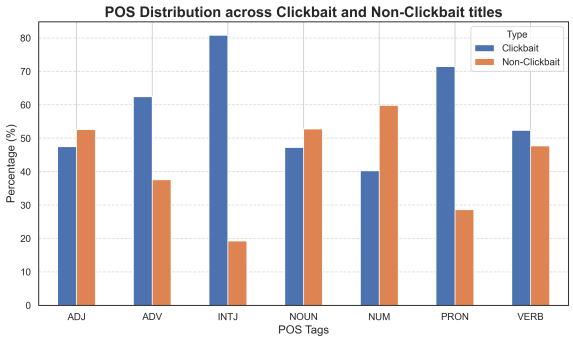

Figure 1: POS Tag Distribution

These differences in linguistic patterns suggest that clickbait headlines aim to capture attention

through more emotional, engaging, and action-driven language, whereas non-clickbait headlines tend to be more neutral and fact-based.

The source-wise distribution from table 3 reinforces the idea that different outlets have varying content strategies, possibly reflecting different business models. For instance, sources like PlayTech may rely more heavily on ad-based revenue models, which encourage the use of clickbait to drive traffic, while PlayTech could focus on a subscription or credibility-based model, prioritizing non-clickbait content.

Digi24 and Go4IT have the lowest proportion of clickbait, suggesting that these sources might focus more on traditional or factual journalism, with less emphasis on sensationalism.

Source	Clickbaits (%)
Descopera	41.53
Digi24	28.43
Go4IT	26.03
PlayTech	73.12

Table 3: Source-wise clickbait distribution

3 Methodology

In this section we discuss the experiments performed for automatic detection of clickbait news based on our introduced dataset. We experiment with various types of supervised machine learning models in order to learn to predict clickbait, including classical ML models, deep learning, and pre-trained transformers.

3.1 Feature Engineering

For some of the models we employ handcrafted features extracted from the news articles. We extract different linguistic stylometric features which might capture the specific style used in clickbait news. At this level, we are interested in capturing stylistic aspects of the news, since clickbait is a general phenomenon in news, which occurs across topics and domains. Unlike in fake news detection, where factuality plays a major role, for clickbait detection, the semantic content of the news is arguably less relevant than the style in which the news is presented. We dedicate a set of experiments to evaluating this hypothesis, by employing various linguistic and stylometric features as input to our ML models.

The first set of linguistic features is related to quantitative aspects of the text, some of which are traditionally used in authorship attribution to capture the style of an author, such as punctuation, part-of-speech distribution, to which we add clickbait-specific features such as the presence of specific keywords, or superlatives.

A second set of features are based on several more complex metrics which have been introduced in previous studies and traditionally used in characterizing a text stylistically from different perspectives, such as: formality score, pronominalisation index, Trager coefficient, readiness to action, aggressiveness coefficient, Coleman-Lieu score, RIX & LIX score.

The full list of handcrafted features is listed in Table A.1 in the Appendix.

3.2 Classification Experiments

We experiment with various kinds of machine learning models, trained on 80% of the dataset and validating the model on the remaining 20%. The first set of experiments use classical ML models applied on 3 different sets of features:

1. TF-IDF features extracted solely from headlines (titles)
2. Numerical linguistic features (detailed in Section 3.1) extracted from headlines
3. Combinations of headline TF-IDF features and numerical linguistic features

We then experiment with deep learning models, including fully-connected feed forward neural networks and LSTM networks. Finally, we use pre-trained transformer models.

3.3 Classical Machine Learning Algorithms

As feature extraction, we compute vectorial representation of the headlines using TF-IDF scores (Ramos, 2003) with 5,000 features based on word unigrams. We combine these with the linguistic features described previously. We compare 6 standard ML algorithms: Logistic Regression, Light Gradient Boosting Machines, XGBoost, Random Forest, Linear Support Vector Machines trained with stochastic gradient descent optimizers as well as passive aggressive algorithms (Crammer et al., 2006). After the experiments, we take the best-performing model and conduct an ablation study in which we explore how each linguistic feature contributed to the performance of the model.

3.4 Deep Learning models

In a second set of experiments, we compare different architectures of neural networks trained from scratch on our dataset. We first experiment with simple fully-connected neural networks with 3 layers and TF-IDF features extracted from the headlines (3 dense layers of 512, 256, and 128 neurons respectively). Secondly, we also train Long-Short Term Memory (LSTM) models with 128 neurons for the LSTM layer, and a dense layer of 64 neurons and a dropout rate of 0.1, using word sequence features with all parameters trained at the same time on our dataset ("vanilla" LSTM); as well as an LSTM model with 128 neurons using as features Word2Vec embeddings trained previously on our data (with an embedding size of 300). All of the neural network models were controlled for overfitting using early stopping. We train the deep learning models on a training set of 80% of the data and evaluate them on the remaining 20% test articles.

3.5 Pre-trained Transformer Models

We finally experiment with pre-trained transformer models, specifically masked language models fine-tuned for text classification. We compare all existing variations of general domain BERT models pre-trained for the Romanian language: BERT-base Romanian (cased & uncased variants) - the first pure BERT models for the Romanian language released in 2020 (Dumitrescu et al., 2020); DistillBERT for Romanian (cased); RoBERT small and RoBERT (Masala et al., 2020). Aside from straightforward fine-tuning of these models for classification on our task, we additionally use a training technique combining the RoBERT transformers (the best-performing ones in the simple setting) with a data augmentation technique (MixUP (Zhang et al., 2018)).

For fine-tuning all transformer models we use a batch size of 32, using weighted sampling in order to compensate for any class imbalance; the models are optimized with the AdamW optimizer and a learning rate of $2e-5$ for a maximum 20 epochs. We used early stopping with a patience of 4 epochs.

The **MixUP** approach, proposed Zhang et al. (2018), is a recent data augmentation technique, used recently especially in computer vision tasks, and less explored for text classification. This approach is based on synthesizing a new sample of points by combining two existing data points. The best-performing transformer model (RoBERT) was fine-tuned by using this new data augmentation technique. Formally, the mixup-transformer is building virtual hidden representations dynamically during the training process (Zhang et al., 2018).

In this setting we first train the transformer model for 5 epochs, with a learning rate of $2*10^{-5}$, using as a training dataset the mix-up augmentation of the original training headlines dataset with variations of the λ parameter used in the MixUP technique to create new datapoints: $\lambda \in \{0.1; 0.3; 0.5; 0.7; 0.9\}$, where λ is in the [0, 1] range, sampled from the Beta distribution and helps to create the combination between two existing data points. In the second phase, we use early stopping with a patience of 2 epochs.

4 Results

This section relates the obtained results in all the different settings discussed in the previous section. All reported results were obtained using single runs. The main metric used for comparison was F1-score. Results on additional metrics can be found in the Appendix.

4.1 Classical Machine Learning algorithms

Table 4 reports the results for the various feature combinations. The best model overall seems to be the Light Gradient Boosting Machine.

Generally, the performance seems to be lowest using only linguistic features. We can infer that the raw text representations (capturing the semantic content of the headline) is still useful to predict clickbait for sci-tech news. We can notice that the scores do not exceed 75% using only linguistic features, Random Forest and Light GBM bringing the best results.

Overall, the combination of both types of features leads to the best performances across models. In the combined setting, Light GBM had the best performance, providing an F1 score of 85.67%, followed by the SVM trained with Stochastic Gradient Descent.

4.1.1 Ablation Study

In order to evaluate the contribution of different linguistic features to the automatic detection of clickbait, we perform ablation studies using our best machine learning model, the LightGBM (LGBMClassifier). We group the linguistic features used based on various linguistic and stylistic aspects. The final features groups are: Grammar features (different

Model	TF-IDF	Linguistic features	TF-IDF + linguistic features
	F1 score	F1 score	F1 score
Logistic Regression	0.8413	0.7220	0.8381
Light GBM	**0.8469**	0.7371	**0.8567**
XGBoost	0.7044	0.6864	0.7105
Random Forest	**0.8464**	**0.7412**	0.8263
SVM + SGD	0.8421	0.7113	**0.8302**
SVM + Passive-Aggr.	0.7659	0.6058	0.7889

Table 4: Results obtained with classical machine learning models with train-test split and different feature settings (TF-IDF or linguistic features extracted from headlines). The best results for each feature setting are marked in bold.

part-of-speech distribution statistics), Readability features, Style features (different linguistic markers of style or particular types of content, such as punctuation, or the presence of numeric data), and Complex features (separate linguistic features such as formality or aggressiveness). The complete mapping of features to their groups is listed in the Appendix. To assess the importance of each group, we measured the model's F1 score on subgroups of features, along with the TF-IDF text representations. Furthermore, for Complex features and Style features, we conducted an additional evaluation by testing each individual feature from these groups in combination with TF-IDF features to measure its independent contribution.

Feature Group-Wise Evaluation The ablation results for each of the four primary categories are as follows:

- Grammar features, comprising features such as the number of words, the length of words, noun usage, and verb types, achieved an F1 score of 0.8516. This suggests that grammatical structure has a strong impact on the model's predictive capability.

- Readability features, which included readability indices such as the RIX, LIX, and Coleman-Liau scores, led to an F1 score of 0.8482. While contributing meaningfully, readability alone did not outperform other feature sets.

- Style features, such as the presence of question marks, exclamation marks, superlatives, and acronyms, achieved the highest group-level F1 score at 0.8531. This highlights the importance of stylistic elements in model performance, suggesting that how information is presented stylistically is crucial for classification tasks.

- Complex features, which measured deeper aspects like f-measure score, trager coefficient, and aggressiveness coefficient, resulted in an F1 score of 0.8487, indicating moderate but useful contributions to model performance.

Individual Feature Ablation in Complex and Style Features We further break down the Complex group of features, which are the most diverse, in order to assess their importance individually through ablation. Additionally, we do the same for the Style features, which obtained the best result in the group-level ablation study.

- For the complex features, the formality score yielded an F1 score of 0.8466, confirming its utility in reflecting intricate behavioral patterns. The Trager coefficient and Aggressiveness coefficient showed F1 scores of 0.8459 and 0.8448, respectively, indicating moderate contributions. The readiness to action coefficient and activity index provided slightly better F1 scores of 0.8472 and 0.8478, respectively, showing their relevance in gauging decision-making readiness in the text.

- For the style features, the presence of question marks achieved the highest F1 score in this group at 0.8553, showing that the use of question marks is a particularly strong stylistic indicator for the clickbait category. The presence of numbers also produced a relatively high F1 score of 0.8501, indicating that numerical references contribute to accurate predictions. Other features like the presence of money terms, acronyms, and terms related to video provided slightly lower F1 scores ranging from 0.8463 to 0.8495, reflecting moderate importance for these stylistic elements.

The full list of feature groups and of the experiments conducted is listed in Table A.2 and table A.3 in the Appendix.

As a results of the ablation style features emerged as the most impactful, particularly elements like the presence of question marks and numerical values. The individual analysis of features within the Complex and Style categories further supported this finding, as stylistic cues such as question marks (indicative of interrogative sentences) proved highly influential for clickbait detection. While grammatical and complex behavioral metrics contributed meaningfully, they did not outperform the stylistic features. This analysis underscores the importance of incorporating a diverse range of features to capture both surface-level presentation and deeper, behavioral insights in textual data classification tasks.

4.2 Deep Learning models

Table 5 reports the results using the neural network based models. We notice that the LSTM models perform better, with the version of vanilla LSTM (using Word2Vec for training embeddings from scratch along with the other parameters) obtaining the best result of 82.95% on the test dataset. Interestingly, these models do not outperform the best classical ML models, which obtain better results than the deep learning models, especially in the combined setting with TF-IDF and linguistic features. This result seems to confirm that linguistic features that capture the stylistic aspects of headlines provide useful information for detecting clickbait in addition to the semantic content.

Model	F1 score
3-Layer NNs	0.7824
LSTM (vanilla)	0.7979
LSTM + Word2Vec	**0.8295**

Table 5: Results obtained by the deep learning models on the test dataset.

4.3 Transformer-based models

The results of transformer based models are shown in Table 6. The RoBERT model obtained the highest F1 score, reaching 88%. Slight improvements were obtained by applying the MixUP technique. These exceed all results obtained with previous models and feature combinations, suggesting the large quantity of data used in pre-training these models is still the most useful for capturing clickbait accurately. Nevertheless, the results obtained using simpler classical ML models with handcrafted features shows that clickbait detection is feasible with reasonable performance even with cheaper solutions.

Model	F1 score
BERT base cased	0.87
BERT base uncased	0.87
DistilBERT base	0.84
RoBERT small	0.88
RoBERT base	**0.88**
RoBERT base + MixUP ($\lambda = 0.1$)	**0.8881**
RoBERT base + MixUP ($\lambda = 0.3$)	0.8837
RoBERT base + MixUP ($\lambda = 0.5$)	0.8836
RoBERT base + MixUP ($\lambda = 0.7$)	0.8837
RoBERT base + MixUP ($\lambda = 0.9$)	0.8836

Table 6: Results on different BERT models on the test dataset

5 Discussion on Performance on Out-of-domain News Data

We finally explore the generalization power of our best-performing model on out-of-domain news by evaluating it on news that is not necessarily part of the scientific-technological sphere.

5.1 Small Set Out-of-domain News

In this experiment, we create a new small dataset of out-of-domain news by manually annotating 50 news items (from politics, external politics, economy, sports, and entertainment categories), using the same annotators and instructions as for our main dataset. Some of these clickbait news was sourced from PlayTech (from a section dedicated to sports and different than technology and science) which is one of the websites used for SciTechBaitRO known to contain clickbait news, whereas the non-clickbait news was taken from Cinemagia[7] (for the movies category) and Digi24 (for politics, economy, and sports).

We test the best RoBERT transformer model and predict the label of our sample of 50 news. We obtain an F1 score of 93.88%. The confusion matrix with the results obtained is shown in Table 8.

[7] https://cinemagia.ro

		Predicted label	
		Non-Clickbait	Clickbait
Actual	Non-Clickbait	24	0
	Clickbait	3	23

Table 7: Confusion Matrix on Out-of-Domain (sample of 50 annotated data) news

Our best model managed to predict the labels on other kinds of news (not only on those belonging to the scientific-technological sphere) with very good performance. Even though the sample is too small to draw definitive conclusions, the results indicate the model seems to generalize quite well on data from new fields and specialties. This suggests that some of the patterns that make news belong to the clickbait category might be universal across topics.

5.2 External Clickbat Dataset - RoCliCO

In a second experiment, we evaluate our best-performing model on an external dataset, the very recently published Romanian news clickbait dataset RoCliCO ((Broscoteanu and Ionescu, 2023)). The authors make available the splits used in their experiments, so we evaluate our models on their test split, to facilitate a direct comparison.

		Predicted label	
		Non-Clickbait	Clickbait
Actual	Non-Clickbait	948	118
	Clickbait	52	389

Table 8: Confusion matrix for our trained model evaluated on the test sample of RoCliCO. (Broscoteanu and Ionescu, 2023)

We can notice in Tables 8 and 9 that performance is remarkably good on this external dataset. We report both F1-score and precision and recall, in order to better understand whether the model has more difficulties with either of the two classes. Our best sci-tech model is better at identifying non-clickbait headlines, with higher precision, recall, and F1-score for class 0 (non-clickbait). The model seems to struggle more with clickbait headlines, with lower precision and F1-score, but better recall. The macro-averaged F1-score obtained with our model trained on SciTechBaitRO corpus and evaluated on RoCliCO is 88.7%, compared to 91.99% obtained by Broscoteanu and Ionescu (2023), with a contrastive learning model trained on the same dataset RoCliCO.

While our best-performing model was specifically designed for science and technology content,

	Precision	Recall	F1-score
Non-Clickbait	0.948	0.889	0.918
Clickbait	0.767	0.882	0.821
Accuracy	0.887	0.887	0.887
Macro Average	0.858	0.887	0.869

Table 9: Classification Report on RoCliCo test dataset

it has still demonstrated a very good performance when applied to the more general task of detecting clickbait. The results show that, even outside its primary domain, the model is capable of identifying clickbait and non-clickbait headlines with good results, making it a useful tool even in different contexts. However, the lower precision for clickbait suggests there could be room for improvement if the model was fine-tuned for specific domains.

6 Conclusions and Further Work

The main goal of this research is to introduce SciTechBaitRO, a new annotated corpus of clickbait news in Romanian, a low-resource language, with a novel focus on the science and technology area., and to evaluate the feasibility of automatically detecting clickbait on these data. We experiment with various machine learning models and features in order to automatically detect the clickbait news, obtaining results upward of 89% F1-score.

We additionally show that classifiers trained on our dataset can perform well on other examples of Romanian online news from different domains as well (i.e. sports, politics, economics, RoCliCo corpus), showing the models are able to generalize to other domains.

While our results show that clickbait can be accurately detected with our methods, some future research could improve performance. Incorporating the body of the news articles as well as using other models (such as Hierarchical Attention Networks (HAN) (Yang et al., 2016) or SetFit) for clickbait news classification could be promising directions.

Limitations

The primary limitation of our clickbait detection model lies in its linguistic scope, being specifically tailored to the Romanian language. While this specialization allows for a nuanced understanding of language-specific features, it also constrains the model's applicability beyond this linguistic context. The linguistic and cultural nuances that are

crucial in identifying clickbait may vary significantly across languages, and as such, further research would be required to adapt and validate the model in different linguistic settings.

Another limitation of this study is the use of ChatGPT as an annotator for labeling the data. While ChatGPT is a highly advanced language model, its performance in labeling can be somewhat inconsistent due to its lack of human judgment and nuanced understanding of context in some instances. The model is trained on a wide variety of data and lacks the cultural and contextual specificity that human annotators possess, which may result in occasional misclassifications. Although majority voting and cross-checking were employed in this process, future work could benefit from more refined or hybrid approaches to improve the reliability of automated annotation.

Lastly, the evolving nature of clickbait tactics presents a challenge to the model's long-term relevance. As strategies for creating clickbait evolve, so too must detection methods. The corpus and model presented here are reflective of the current state of clickbait in Romanian science and technology news, and ongoing updates may be necessary to maintain accuracy over time.

Ethics Statement

The primary goal of our study is to contribute to the responsible dissemination of information by developing tools that can help mitigate the spread of clickbait, which often misleads or manipulates readers.

The data used in this study was collected from publicly available news articles, ensuring that no private or sensitive information was compromised. The news articles are freely accessible to the public without any type of subscription. we adhere to European regulations that permit researchers to use publicly available data on the web for non-commercial research purposes. Specifically, our data usage aligns with Directive (EU) 2019/790 of the European Parliament and the Council on copyright and related rights in the Digital Single Market [8]. In accordance with these regulations, we release our corpus as open-source under a non-commercial share-alike license agreement, ensuring that the dataset remains available for further research and development under the same terms.

We have taken care to anonymize the sources of data where necessary and to avoid any potential bias in the selection and labeling of the data. Additionally, we acknowledge that some news samples in our corpus may reference public figures or other identifiable individuals. Should we receive a request to anonymize such references, we will promptly and respectfully comply, ensuring that the privacy and rights of individuals are upheld.

Our focus on Romanian science and technology news reflects an effort to address clickbait within a specific, manageable scope, while acknowledging that these findings may not be directly transferable to other languages or domains without further research.

We are aware that clickbait detection models can have significant implications for media, journalism, and public information. Therefore, we have approached the development of our model with caution, aiming to minimize false positives that could unjustly flag legitimate news content. Our research is intended to support, not undermine, the journalistic process by providing tools that enhance the quality of information reaching the public.

Finally, we recognize the importance of transparency in our research. All methodologies, data sets, and results are fully documented to allow for reproducibility and further scrutiny by the research community. We invite feedback and collaboration to refine and improve upon this work, with the ultimate aim of promoting a more informed and discerning public discourse.

We do not employ any AI Assistants in the writing of this study.

[8] https://eur-lex.europa.eu/eli/dir/2019/790/oj

Acknowledgment

This research is partially supported by the POCIDIF project in Action 1.2 "Romanian Hub for Artificial Intelli- gence" MySMIS no. 310483".

References

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *The Journal of Reading*, 26.

The Editors of Encyclopaedia. Britannica. 2024. "yellow journalism". Accessed 16 August 2024.

Daria-Mihaela Broscoteanu and Radu Tudor Ionescu. 2023. A novel contrastive learning method for clickbait detection on roclico: A romanian clickbait corpus of news articles.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait:

Detecting and preventing clickbaits in online news media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Manfred K Warmuth. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(3).

Edgar Dale and Jeanne Chall. Dale-chall easy word list. http://countwordsworth.com/download/DaleChallEasyWordList.txt6.Davis,J.,Goadrich,M.:Therelationship.

Stefan Daniel Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. In *FINDINGS*.

Bhanuka Gamage, Adnan Labib, Aisha Joomun, Chern Hong Lim, and KokSheik Wong. 2021. Baitradar: A multi-model clickbait detection algorithm using deep learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2665–2669. IEEE.

Ayse Geckil, Ahmet Anil Mungen, Esra Gündogan, and Mehmet Kaya. 2018. A clickbait detection method on news sites. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 932–937.

Sura Genç and Elif Surer. 2021. Clickbaittr: Dataset for clickbait detection from turkish news sites and social media with a comparative analysis via machine learning algorithms. *Journal of Information Science*, page 016555152110077.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20:1 – 18.

Francis Heylighen, Jean-Marc Dewaele, and Léo Apostel. 1999. Formality of language: definition, measurement and behavioral determinants.

Mini Jain, Peya Mowar, Ruchika Goel, and Dinesh K Vishwakarma. 2021. Clickbait in social media: detection and analysis of the bait. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Tatiana A. Litvinova, Olga Zagorovskaya, Olga Litvinova, and Pavel Seredin. 2016. Profiling a set of personality traits of a text's author: A corpus-based approach. In *SPECOM*.

Mihai Masala, Stefan Ruseti, and M. Dascalu. 2020. Robert – a romanian bert model. In *COLING*.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig MacDonald, and Christina Lioma. Terrier : A high performance and scalable information retrieval platform.

Aralda Păcurar and Ciprian Oprişa. 2023. Using artificial intelligence to fight clickbait in romanian news articles. In *2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 397–404. IEEE.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *ECIR*.

Juan Enrique Ramos. 2003. Using tf-idf to determine word relevance in document queries.

Nils Reimers and Iryna Gurevych. 2019. 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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Viktoriia Vasyliuk, Yuliia Shyika, and Tetiana Shestakevych. 2020. Information system of psycholinguistic text analysis. In *COLINS*.

Andika William and Yunita Sari. 2020. Click-id: A novel dataset for indonesian clickbait headlines. *Data in Brief*, 32:106231.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412.

A Appendix

A.1 Prompt used for data annotation

The following prompt has been used for the Large Language Model (LLM) to annotate our data.

"Te rog analizeaza urmatoarele titluri de stiri si spune daca sunt clickbait sau nu. Da-mi rezultatul sub forma unei liste de 0 si 1, unde elementul de pe pozitia i corespunde propozitiei de pe randul i, si unde 1 inseamna clickbait si 0 inseamna ca nu este clickbait. Criteriile de analiza sunt urmatoarele: daca titlul contine cuvinte ca "VIDEO", "ciudat", "inspaimantator", "mister", "PHOTO GALERY", "periculos" sau sinonime care tind sa dramatizeze, sau daca titlul pune o intrebare (Ce, Cum, Cine,

Unde, Cat, Cui). Titlurile sunt mai jos, cate unul pe fiecare rand."

A.2 Handcrafted features

Table A.1 displays all of the linguistic-based features created based on the headlines and that were used in the classification models.

A.3 Ablation Study

Table A.2 displays the feature groups created and their corresponding features. The best-performing machine learning model, Light GBM, was tested on these groups and on individual features. The results of these experiments are displayed in table A.3.

A.4 Infrastructure and Configuration

In order to experiment, but also to train various models and try different approaches, we used 3 different work environments and we paralleled the work in several directions using those 3 hardware environments.

- NVIDIA GeForce GTX 1050 Ti with 4GB memory
- NVIDIA Tesla P100 with 16GB memory provided by Kaggle[9]
- NVIDIA Tesla K80 with 12GB GPU memory provided by Google Colab[10]

The training times and number of epochs for the transformer models are reported in Table A.4.

A.5 Libraries

The libraries used for data processing and machine learning:

- $nltk3.8$ - feature extraction (TF-IDF features and tokenization for the classical machine learning models), POS-tagging
- $sklearn1.5.1$ - classical machine learning models training and evaluation
- $gensim4.0$ - training Word2Vec embeddings
- $Keras3.5.0$, $Tensorflow2.16$, $pytorch$ and $transformers$ for deep learning and transformer models training and evaluation
-

[9] https://www.kaggle.com/docs/efficient-gpu-usage

[10] https://colab.research.google.com/

Feature	Type	Description
no_of_words	numeric	number of words
avg_words_length	numeric	average words length
no_of_common_nouns	numeric	number of common nouns
no_of_proper_nouns	numeric	number of proper nouns
no_of_adjectives	numeric	number of adjectives
no_of_2nd_person_verbs	numeric	number of verbs at the 2^{nd} person
no_of_3rd_person_verbs	numeric	number of verbs at the 3^{rd} person
no_of_verbs	numeric	total number of verbs
no_of_demonstrative_pronouns	numeric	number of demonstrative pronouns
no_of_personal_pronouns	numeric	number of personal pronouns
f_measure_score	numeric	formality score as stated in (Heylighen et al., 1999)
pronominalisation	numeric	pronominalisation index as stated in (Litvinova et al., 2016)
trager_coefficient	numeric	Trager coefficient as stated in (Litvinova et al., 2016)
aggressiveness_coefficient	numeric	aggressiveness coefficient as stated in (Vasyliuk et al., 2020)
readiness_to_action_coefficient	numeric	directness coefficient as stated in (Vasyliuk et al., 2020)
activity_index	numeric	activity index as stated in (Vasyliuk et al., 2020)
RIX_score	numeric	readability score as stated in (Anderson, 1983)
LIX_score	numeric	readability score as stated in (Anderson, 1983)
CL_score	numeric	Coleman-Liau score as stated in (Coleman and Liau, 1975)
superlatives	boolean	whether a headline contains superlatives
no_of_qm	numeric	number of question marks
qm_present	boolean	whether a headline contains question marks
no_of_em	numeric	number of exclamation marks
em_present	boolean	whether a headline contains exclamation marks
money_present	boolean	whether a headline contains different currencies (RON, EURO)
questions_present	boolean	whether a headline contains questions like "Ce, De ce, Cand, Cine, Care, Cum"
acronyms_present	boolean	whether a headline contains acronyms
numbers_present	boolean	whether a headline contains numbers
video_present	boolean	whether a headline contains video or not

Table A.1: Characteristics and features derived from the headlines

Feature Group	Features
Grammar features	no_of_words, avg_words_length, no_of_common_nouns, no_of_proper_nouns, no_of_adjectives, no_of_2nd_person_verbs, no_of_3rd_person_verbs, no_of_verbs, no_of_demonstrative_prons, no_of_personal_prons
Readability features	RIX_score, LIX_score, CL_score
Style features	superlatives, no_of_qm, qm_present, no_of_em, em_present, money_present, questions_present, acronyms_present, numbers_present, video_present
Complex features	f_measure_score, trager_coefficient, aggressiveness_coefficient, readiness_to_action_coefficient, activity_index

Table A.2: Features used for each feature group

Feature Group / Feature	F1 Score
Grammar features	0.8516
Readability features	0.8482
Style features	0.8531
superlatives	0.8470
no_of_qm	0.8553
qm_present	0.8553
no_of_em	0.8470
em_present	0.8470
money_present	0.8470
questions_present	0.8463
acronyms_present	0.8470
numbers_present	0.8501
video_present	0.8495
Complex features	0.8487
f_measure_score	0.8466
trager_coefficient	0.8459
aggressiveness_coefficient	0.8448
readiness_to_action_coefficient	0.8472
activity_index	0.8478

Table A.3: F1 Scores for Different Feature Groups and Individual Features

Algorithm	Epochs	Training time
BERT base cased	5	00:08:28
BERT base uncased	5	00:09:27
DistilBERT base	15	00:15:49
RoBERT small	7	00:05:13
RoBERT base	5	00:09:45
RoBERT base + MixUP ($\lambda = 0.1$)	2	00:14:15
RoBERT base + MixUP ($\lambda = 0.3$)	2	00:13:38
RoBERT base + MixUP ($\lambda = 0.5$)	2	00:13:38
RoBERT base + MixUP ($\lambda = 0.7$)	2	00:14:17
RoBERT base + MixUP ($\lambda = 0.9$)	2	00:15:20

Table A.4: Training times (in hh:mm:ss) and epochs for BERT models.