Computational Humor Detection on Novel Romanian Dataset

Mihalache Diana
mihalachediana15@gmail.com

Ursu Andrei
andreiursu234@gmail.com

Abstract

This paper proposes an in-depth study of a machine learning system and pretrained transformer models aimed at detecting the punchline and classifying the humorous sentences published in Romanian, and also comparing the results with LORA adapted models and the initial ones. A corpus of 2.400 manually annotated data was used, 50% of which contain humor and the rest of them are reflecting natural conversations. We then proceeded with conducting experiments: Dataset analysis, Syntax analysis, POS-tagging, followed by classical machine learning algorithms and transformer models (BERT). The best result is given by BERT models, especially XLM-RoBERTa, with an accuracy score of 98%.

1 Introduction

As stated in the Cambridge Dictionary, humor is defined by "the ability to be amused by something seen, heard, or thought about, sometimes causing you to smile or laugh". Despite this definition, we cannot explain in a universally accepted theory “what, why, how, when and to whom it is funny” (Taylor and Mazlack, 2004). Humor is representing a big part of the human experience, reflecting cultural nuances, social contexts, and cognitive processes. In addition, every known human civilization has also had at least some form of humor to make others laugh (Caron, 2002; Gervais and Wilson, 2005). Despite its ubiquitous presence in daily life, humor remains a complex phenomenon to define and understand, due to its subjective nature and diverse forms.

In recent years, the advancement of natural language processing (NLP) has sparked interest in the automated detection of humor. This task involves the identification and classification of humorous content within text, which is challenging due to the subtleties and variations of humor across different languages and cultures. Detecting humor requires an understanding of linguistic ambiguity, sarcasm, puns, and other rhetorical devices that convey humor. The ability to accurately detect humor has practical applications in several domains, including sentiment analysis, chatbots, content moderation, and entertainment. For example, enhancing chatbot interactions with humor can lead to more engaging and human-like conversations, while content-moderation systems can identify and filter potentially inappropriate or offensive humor in online platforms.

In this article, we explore several computational techniques that have been used for the detection and classification of humor, highlighting the challenges and limitations faced in the field. We introduce a new dataset and examine various models and linguistic features. Our aim is to provide a comprehensive overview of the current state of humor detection and suggest avenues for future research that could improve the effectiveness and accuracy of these systems.

2 Related Work

Recently, the field of humor and satire detection has become an area of interest, as researchers have explored various approaches to understand and automate this process. This NLP task is quite difficult because of its complexity and the wide variety of humor types that must be identified. However, most studies on this topic are conducted using English-language datasets. Given that there are very few Romanian language datasets for NLP tasks, we propose both the continuation of the ***SaRoCo: Detecting Satire in a Novel Romanian Corpus of News Articles*** article (Rogoz et al., 2021) and its extension with a newly created and manually annotated dataset, using it as a baseline. This article proposes a new dataset with 55,608 public news articles. Their results indicate that the machine-level accuracy for detecting satire in Romanian is relatively low, at under 73% on the test set, when compared to the human-level accuracy of 87%. The SaRoCo dataset was used also in ***Adversarial Capsule Networks for Romanian Satire Detection and Sentiment Analysis*** (Echim et al., 2023) which proposes a novel approach for detecting satire and performing sentiment analysis in Romanian texts using Adversarial Capsule Networks (ACNs). The proposed framework surpasses existing methods for both tasks, achieving an accuracy of up to 99.08%, thereby demonstrating a significant improvement.

3 Methodology

We explore several computational techniques that have been used for the detection and classification of humor on a new dataset and examine various models and linguistic features that are presented in this chapter.

3.1 Dataset

The dataset contains 2.400 sentences, divided equally as humor sentences (marked with 1), which also have the punchline extracted, and non-humor sentences (marked with 0). The sentences that contain humor are extracted from stand-up specials, from both male and female comedians. The natural conversation sentences were extracted from Reddit. The data annotation was executed manually from both a male and a female perspective, so that the dataset reflects a balanced viewpoint and minimizes gender bias in the classification and analysis of humor. This dual-perspective approach aims to enhance the accuracy and fairness of the models trained on this data. We shuffled the data before splitting it into 3 datasets: 75% for the training phase of the model, 15% for the validation phase and 15% for the testing phase.

Dataset	Sentences
Training	1680
Validation	360
Testing	360

Table 1: Datasets

3.2 Dataset Analysis

To fully understand our dataset, we opted for an exploratory data analysis to ensure that the data in the dataset is independent and identically distributed. Thus, in the following figures, we explore the data structure using n-grams and POS tagging.

Label	Words	Characters
Humor	18.52	81.09
Punchline	5.53	12.83
Non Humor	15.71	75.13

Table 2: Average Lengths of the labels

Figures 1 and 2 present an analysis of the most frequent nouns found in two different classes of a humor dataset. The two classes represent non-humorous (Class 0) and humorous (Class 1) texts. The differences in vocabulary usage provide insights into the nature of humor in the dataset.

In the non-humorous class, the words suggest a focus on education, work, and general life experiences, indicating a more serious or neutral tone.

In the humorous class we notice a stronger presence of words related to people, relationships, and potentially humorous or taboo topics such as “sex” and “niciodata”. This suggests that humor in this dataset is often derived from personal experiences, social interactions, and unexpected situations.

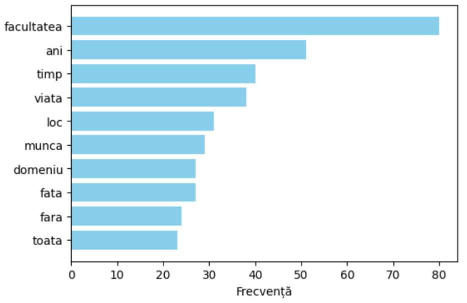

Figure 1: The first 10 most common NOUNS from the Non Humor Label

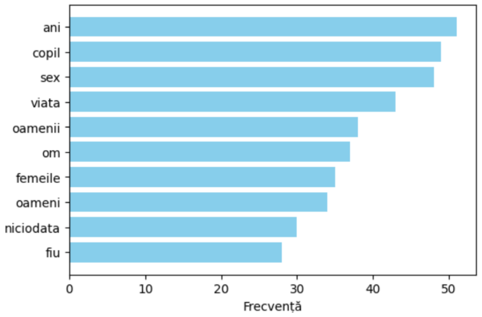

Figure 2: The first 10 most common NOUNS from the Humor Label

Figures 3 and 4 show the distribution of sentence lengths, measured by word count, for humorous and non-humorous texts. The first plot, for humorous texts, reveals a bimodal distribution. There is a peak around 10-12 words per sentence, as well as a larger peak around 100-120 words per sentence. This suggests that humorous writing often utilizes both shorter, punchier sentences as well as longer, more elaborate sentences. The second plot, for non-humorous texts, exhibits a more typical bell-shaped distribution centered around 20-25 words per sentence. This indicates that non-humorous writing tends to have a more consistent sentence structure, without the same variability seen in the humorous texts. The humorous writing leverages a wider range of sentence lengths compared to non-humorous writing, likely to create different rhythms and patterns that contribute to the comedic effect. The bimodal distribution for humorous texts is a notable difference from the more unimodal distribution for non-humorous texts.

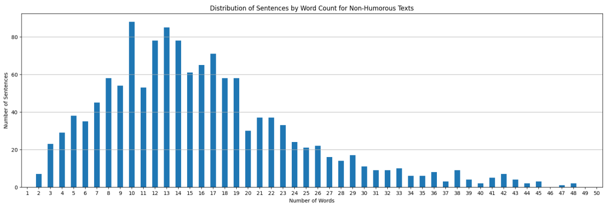

Figure 3: Non-Humor Word Count

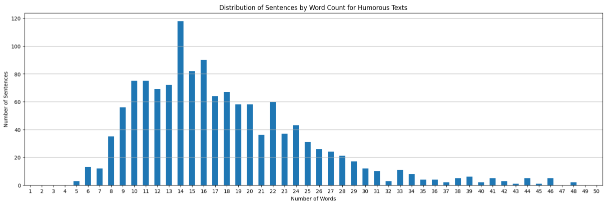

Figure 4: Humor Word Count

3.3 Preprocessing

For these, we preprocessed the texts using the TF-IDF vectorizer, and then used them to train the models. Their performance is evaluated on the validation and evaluation datasets, measuring accuracy using the `classification_report` function.

The preprocessing task for Romanian phrases involves two main steps: replacing special characters and removing prepositions and stop words. The first part is designed to handle the replacement of specific Romanian special characters with their non-accented equivalents. The replacements are defined as follows: `ă` is replaced with `a`, `î` and `â` are both replaced with `i`, `ș` is replaced with `s`, `ț` is replaced with `t`. The second part is used to remove both prepositions and stop words from the text, using a list with common stopwords in Romanian (Gene Diaz, 2016) and the Stanza package (Qi et al., 2020) that includes various functions such as tokenization and part-of-speech (POS) tagging. The stop words list includes common words such as "sau", "in", "pe", "la", "cu", etc., which are typically filtered out to reduce noise in the text data. By removing these elements, the preprocessing aims to retain only the most informative parts of the text, which are then used for further analysis or model training.

3.4 Methods

We used models[1] as follows:

- **RoBERT** (Dumitrescu et al., 2020) is a Romanian language model based on the BERT architecture, designed for tasks like text classification, sentiment analysis, and named entity recognition in Romanian.

- **XLM-RoBERTa (XLM-R)** (Conneau et al., 2020) is a transformer-based multilingual language model that extends RoBERTa to support over 100 languages. It is designed for natural language understanding tasks across diverse languages, using a shared vocabulary and training on large-scale multilingual datasets.

To explore parameter-efficient fine-tuning, we also investigated the effectiveness of Low-Rank Adaptation (LoRA). LoRA freezes the pre-trained weights of the large language model and introduces a small number of new rank-decomposition matrices into each layer of the Transformer architecture. By only training these low-rank matrices, LoRA significantly reduces the number of trainable parameters, leading to more efficient adaptation to downstream tasks. The test accuracy results for both models with and without LoRA are presented in Table 3.

Model	**Test Accuracy**
RoBERT	0.97
RoBERT with LORA	0.96
XLM-RoBERTa	0.98
XLM-RoBERTa with LORA	0.9667

Table 3: Test Accuracy Results

[1] The repository on GitHub with the code.

Based on the results presented in Table 3, we can analyze the performance of RoBERT and XLM-RoBERTa with and without the Low-Rank Adaptation (LoRA) technique for the humor detection task. For RoBERT, the application of LoRA resulted in a slight decrease in test accuracy, from 0.97 without LoRA to 0.96 with LoRA. This suggests that while LoRA offers parameter efficiency, there was a minimal trade-off in performance for this specific model and task. In contrast, XLM-RoBERTa demonstrated a different behavior. Without LoRA, it achieved a test accuracy of 0.98. When fine-tuned with LoRA, the accuracy was 0.9667, indicating a small reduction in performance compared to the full fine-tuning of XLM-RoBERTa. Overall, LoRA showed potential as a parameter-efficient fine-tuning method for both models, with a minor impact on the final accuracy, particularly for RoBERT. The results suggest that the effectiveness of LoRA might be model-dependent, requiring careful consideration of the trade-off between parameter reduction and performance for each specific architecture and task.

The confusion matrices of the models are presented in the appendix in figures: 5, 6, 7, 8.

4 Future Work

We will continue expanding the dataset with more data and strive to diversify it as much as possible. Additionally, we will explore and test more robust methods for detecting the humorous part of each joke.

5 Conclusion

In conclusion, this research proposes a new dataset for detecting humor in medium and short-length sentences and serves as a valid starting point for analyzing and detecting humorous structures in Romanian texts.

Limitations

When it comes to the binary classification task, we noticed that the results were far beyond our expectations. In this context, we can still question how the models learn, specifically whether they truly understand humor or simply recognize certain specific linguistic structures. Although the obtained scores for punchline detection are quite good, we must mention that this model tends to learn more about the location of the punchline within the supporting texts. Thus, if we input a text with a punchline that is much longer than usual or placed in an unusual location, the model tends to identify as the punchline the groups of words located in the 70%-80% segment of the sentence's length.

6 Author contribution statement

We both shared the tasks for this paper as we considered equally. We both designed the research methodology, collected and preprocessed the dataset, ensuring data quality and consistency. Andrei conducted the majority of the experiments and performed the statistical analysis for both of the tasks. Diana trained the transformer models for the classification task, and wrote the majority of the manuscript, structured the results, and revised the paper for clarity. We both reviewed and approved the final version of the manuscript.

References

James E. Caron. 2002. From ethology to aesthetics: Evolution as a theoretical paradigm for research on laughter, humor, and other comic phenomena. *HUMOR*, 15(3):245–281.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Sebastian-Vasile Echim, Răzvan-Alexandru Smădu, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Florin Pop. 2023. *Adversarial Capsule Networks for Romanian Satire Detection and Sentiment Analysis*, page 428–442. Springer Nature Switzerland.

Janimo Jani Monoses Gene Diaz. 2016. Stopwords for romanian (stopwords-ro).

Matthew Gervais and David Sloan Wilson. 2005. The evolution and functions of laughter and humor: A synthetic approach. *The Quarterly Review of Biology*, 80(4):395–430.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Ana-Cristina Rogoz, Mihaela Gaman, and Radu Tudor Ionescu. 2021. Saroco: Detecting satire in a novel romanian corpus of news articles. In *Proceedings of ACL*, page TBD.

J.M. Taylor and L.J. Mazlack. 2004. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

A Appendix

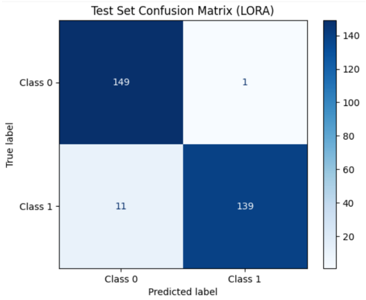

Figure 5: RoBERT Confusion Matrix

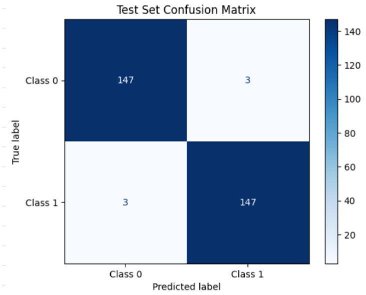

Figure 6: Roberta Confusion Matrix

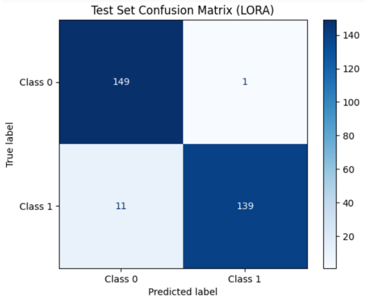

Figure 7: RoBERT with LORA Confusion Matrix

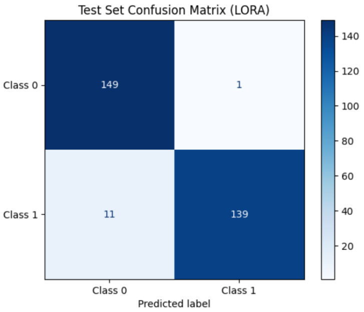

Figure 8: Roberta with LORA Confusion Matrix

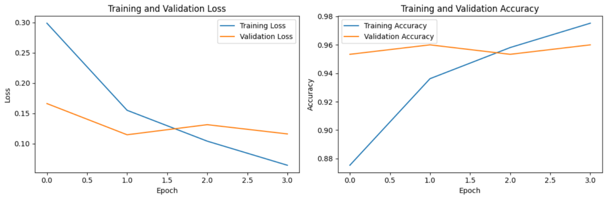

Figure 9: RoBERT training and validation

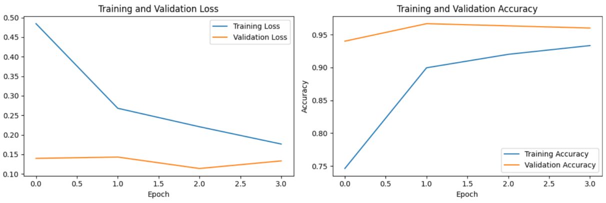

Figure 10: RoBERTa training and validation