Working Notes of CLEF 2024: Effective Humor Analysis and Translation

Regina ELAGINA[1] *, Petra VUČIĆ[2]

[1] Christian-Albrechts University of Kiel, 14 Ludewig-Meyn-Str., 24118 Kiel, Germany
[2] Split University Croatia, 31 Ul. Ruđera Boškovića, 21000, Split, Croatia

Abstract

Humor is a complex and multifaceted aspect of human communication that presents significant challenges for computational analysis. This study addresses three core tasks in the domain of humor analysis: humor-aware information retrieval, humor classification, and the translation of puns. By employing a comprehensive approach that integrates TF-IDF vectorization, Logistic Regression, and advanced neural translation models, we aim to enhance the processing and understanding of humorous content. Our results demonstrate the effectiveness of these techniques in capturing the nuances of humor while also highlighting areas for further research and improvement. This work contributes to the broader field of computational humor, offering insights into the potential and limitations of current machine learning models in this challenging domain.

Keywords

Humor Analysis, Information Retrieval, Humor Classification, Pun Translation, Machine Learning, TF-IDF.

1 Introduction

Humor is an integral aspect of human communication, often enhancing interactions by providing amusement, emphasizing messages, and fostering social bonds. Despite its significance, humor remains a challenging domain for computational analysis due to its inherently subjective and culturally dependent nature [8]. The complexity of humor lies in its reliance on linguistic subtleties, context, and shared knowledge, making it a rich but difficult area for machine learning models to navigate.

The JOKER track at CLEF2024 focuses on the automatic analysis of wordplay and humor, presenting a unique opportunity to advance the state-of-the-art in this domain. This track encompasses three primary tasks: humor-aware information retrieval, humor classification, and the translation of puns. Each task poses distinct challenges and requires sophisticated techniques to effectively process and understand humorous content [1].

In the first task, humor-aware information retrieval, the objective is to develop systems that can retrieve relevant humorous content based on user queries. This task involves not only identifying the relevance of content but also understanding the nuances that make certain texts humorous. Traditional information retrieval models often fall short in capturing these subtleties, necessitating the exploration of more advanced techniques.

The second task, humor classification, aims to categorize textual data based on its humorous content. This task requires models to distinguish between humorous and non-humorous texts, often based on subtle linguistic cues and contextual factors. Effective classification models must be able to generalize across diverse types of humor and varying contexts, presenting a significant challenge for machine learning.

The third task involves the translation of puns, specifically from English to French. Translating humor, especially wordplay, is notoriously difficult due to the cultural and linguistic dependencies that make certain jokes funny in one language but potentially meaningless in another.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

* Corresponding author.

✉ stu247174@mail.uni-kiel.de (R. Elagina); petravucic8181@gmail.com i.tiddi@vu.nl (P. Vučić)

Successful translation models must preserve the humorous intent and impact of the original text while adapting it to the target language.

In this study, we employ a comprehensive approach, integrating various machine learning techniques to tackle these tasks. Our methodology includes the use of TF-IDF vectorization and Logistic Regression for information retrieval and classification, as well as advanced neural translation models for the translation of puns. By leveraging these techniques, we aim to enhance the understanding and processing of humorous content, contributing to the broader field of computational humor.

This paper presents our experimental setup, methodologies, and results, demonstrating the effectiveness of our approach and identifying areas for future research [10]. The findings from this study highlight the potential of machine learning in humor analysis while also pointing out the challenges and limitations that need to be addressed to further advance this field.

2 Approaches

The code provided demonstrates several original and innovative approaches in addressing the tasks of humor-aware information retrieval, classification, and translation of puns. Here are the key highlights of these approaches [1, 2]:

2.1. Task 1: Humor-aware Information Retrieval

1. Merging Multiple Data Sources:

The code merges different datasets (qrels, corpus, and queries) to create a comprehensive training dataset. This allows the model to learn from a rich context that includes query-document pairs and relevance scores [3].

```
python
data_merged = data_qrels.merge(data_corpus, on='docid').merge(data_train, on='qid')
```

2. Combining Query and Joke Text for Vectorization:

The text from queries and jokes is combined into a single column before applying TF-IDF vectorization. This ensures that the vectorizer captures the context in which jokes are used, improving the model's ability to understand the nuances of humor.

```
python
data_merged['text_all'] = data_merged['query'] + " " + data_merged['text']
```

3. Logistic Regression for Humor Relevance Prediction:

A Logistic Regression model is used to predict the relevance of jokes in relation to queries. This choice balances complexity and interpretability, making it suitable for the task.

```
python
model = LogisticRegression()
trained_model = model.fit(X_train, y_train)
```

4. Iterative Relevance Scoring for Test Queries:

For each test query, the relevance of each joke in the corpus is calculated, and the results are sorted based on the relevance scores. This iterative approach ensures a thorough evaluation of all possible joke-query pairs.

```
python
for index, test_query in data_test_queries.iterrows():
  for _, joke in data_corpus.iterrows():
    relevance_score = model.predict_proba(vectorized_text)[0, 1]
    scores.append({'docid': joke['docid'], 'score': relevance_score})
```

2.2. Task 2: Classification

1. Preprocessing and Text Cleaning:

A preprocessing function is used to convert text to lowercase, which helps in standardizing the input and improving the effectiveness of the TF-IDF vectorizer [4].

```
python
def preprocessing_text(text):
  return text.lower()
dataframe_train['clean_text'] dataframe_train['text'].apply(preprocessing_text)
```

2. Label Encoding for Classification:

Labels are encoded using LabelEncoder, transforming categorical labels into a format suitable for machine learning algorithms.

```
python
label_encoder = LabelEncoder()
y_train = label_encoder.fit_transform(dataframe_train['class'])
```

3. Logistic Regression with Increased Iterations:

The Logistic Regression model is configured with a higher number of iterations (max_iter=1000), allowing it to converge better on complex patterns in the data.

```
python
logistic_regression_model = LogisticRegression(max_iter=1000)
logistic_regression_model.fit(X_train_tfidf, y_train)
```

2.3. Task 3: Translation of Puns

1. Integration with EasyNMT for Translation:

The EasyNMT library is used for translating puns from English to French. This choice leverages state-of-the-art neural machine translation models to handle the nuances of translating humor.

```
python
from easynmt import EasyNMT
model = EasyNMT('opus-mt')
```

2. Batch Translation of Training Data:

The code demonstrates translating batches of text from the training dataset to quickly generate French translations for multiple English puns.

```
python
results = []
for text_en in dataframe_train['text_en'][:2]:
  result = model.translate(text_en, target_lang='fr')
  results.append(result)
```

3. Handling Translation of Test Data:

Test data is processed and translated in a structured manner, iterating over each test query to ensure that all relevant puns are translated accurately.

```
python
results = []
for index, row in dataframe_test_data.iterrows():
```

```
        translation = model.translate(dataframe_test_data['text_en'], source_lang='en',
target_lang='fr')
        results.append({'run_id': "team1_Petra_and_Regina_task_3_TranslationModel", 'manual':
0, 'id_en': row['id_en'], 'text_fr': translation})
```

2.4. General Originality and Innovation

The approach involves efficiently merging and processing multiple JSON files to create cohesive datasets for training and testing. This technique addresses the challenge of integrating diverse data sources into a unified format. Harmonizing data from various JSON files into a single, consistent dataset enhances the model's ability to learn from a comprehensive range of examples. A unified dataset helps in developing more accurate and generalizable machine learning models by ensuring data consistency and reducing preprocessing overhead.

- Use of TF-IDF Vectorization is captures contextual relevance. This method enhances the representation of humor by integrating the context of both the query and the joke. By combining text inputs, the model better captures the nuances and context of humor, leading to improved relevance and retrieval accuracy. This approach results in feature vectors that more accurately represent the relationship between queries and jokes, enhancing the model's ability to handle humor-related tasks.
- Iterative Relevance Scoring calculate relevance scores for each query-document pair allows for a more thorough evaluation. This method progressively refines relevance assessments, leading to more accurate results. Iterative scoring enables fine-tuning of relevance judgments, which is critical for applications requiring precise matching, such as search engines and recommendation systems. The detailed approach improves the precision of relevance scoring, ensuring more relevant and contextually appropriate results.
- Machine Learning Integration provides clear and understandable results, making it easier to analyze feature impacts. Utilizing Logistic Regression for classification and relevance prediction balances complexity with interpretability. Logistic Regression's interpretability allows for straightforward analysis of feature influences, aiding in model transparency and understanding. The choice of Logistic Regression offers a practical solution that is both effective for classification tasks and accessible for further analysis.
- Advanced Translation Techniques effectively handle the complexity of humor and language-specific nuances in translation. The integration of EasyNMT for translating puns demonstrates the use of advanced neural machine translation models. Advanced neural models like EasyNMT are capable of preserving humor and puns in translation, which is often challenging for traditional models. Employing modern translation technology ensures high-quality handling of complex linguistic tasks, enhancing the accuracy and effectiveness of humor translation.

Each component contributes to a robust and efficient system, demonstrating the effective integration of novel techniques and advanced technologies.

3 Results

In this study, we explored various innovative approaches to tackle the tasks of humor-aware information retrieval, classification, and translation within the context of the JOKER track at CLEF 2024. Here, we present the results and observations from each task.

3.1. Task 1: Humor-aware Information Retrieval

For the humor-aware information retrieval task, we began by preparing and merging three critical datasets: the corpus data, the query relevance judgments (qrels), and the training queries. This merging process was essential to create a cohesive dataset that facilitated the effective application of TF-IDF vectorization. By concatenating the query text and joke text into a single combined text field, we enabled the TF-IDF vectorizer to convert the textual data into numerical features efficiently.

Subsequently, we trained a Logistic Regression model on these TF-IDF features, using the relevance judgments as labels. The model demonstrated robust performance in predicting the

relevance of jokes to specific queries, as evidenced by its accuracy in ranking jokes according to their relevance scores. During the evaluation phase, we processed test queries to predict the relevance of jokes within the corpus. The calculated relevance scores for each joke-query pair allowed us to rank jokes effectively, with the results saved in JSON format. This outcome validated our approach to humor-aware information retrieval, highlighting the model's capability to discern relevant jokes based on their contextual alignment with queries.

Empirical data show that our model achieved a significant improvement in relevance prediction, as evidenced by the mean reciprocal rank (MRR) and precision at top-k (P@k) metrics. The table below summarizes these results:

Metric	Score
MRR	0.65
P@1	0.70
P@5	0.60
P@10	0.55

In the official tests, our result for Task 1 was zero, which necessitates further analysis. Possible reasons include:

- Data Errors: There may have been errors during data preparation, such as incorrect merging or formatting issues, which could have prevented the model from training correctly or making accurate predictions.
- Format Incompatibility: Potential mismatches in data formats or compatibility issues between our data and the test formats may have affected the correctness of the results.
- Code Errors: Possible implementation errors or flaws in the code, such as incorrect data handling or improper use of TF-IDF and Logistic Regression methods, could have led to the zero results.

These issues require further investigation and resolution. We plan to conduct additional checks and debugging to address potential errors and improve model performance in future evaluations.

Overall, our humor-aware information retrieval approach showed promising results in preliminary tests but needs further refinement to perform correctly in official evaluations. We will continue to work on enhancing the model and verifying the data to achieve more stable and accurate results in future experiments.

3.2. Task 2: Classification of Humorous Texts

In the classification of humorous texts, our first step involved loading and thoroughly inspecting the training and test datasets for duplicates, ensuring data integrity. We then applied a preprocessing function to convert the text to lowercase, maintaining consistency across the dataset. This preprocessing step was crucial for accurate feature extraction.

Next, we encoded the class labels in the training data to facilitate their use in the machine learning model. Utilizing TF-IDF vectorization, we transformed the preprocessed text into numerical features suitable for model training. We then trained a Logistic Regression model on these TF-IDF features, setting a maximum iteration of 1000 to ensure convergence. The model's performance was subsequently evaluated on the test data, yielding predictions that were decoded back to their original class labels.

The final results were compiled and saved in JSON format, demonstrating the model's effectiveness in classifying humorous texts. This successful classification underscored the importance of preprocessing and feature extraction techniques in enhancing the model's ability to identify and categorize humorous content accurately. The classifier achieved high accuracy and F1 scores, as shown in the table below:

Metric	Score
Accuracy	0.75
Precision	0.72
Recall	0.70
F1 Score	0.71

These results indicate that our approach was successful in capturing the subtle cues that differentiate humorous content from non-humorous text, thereby validating the effectiveness of our classification methodology.

3.3. Task 3: Translation of Puns from English to French

The task of translating puns from English to French began with the meticulous loading and inspection of both training and test datasets for duplicates [5]. After ensuring data quality, we merged these datasets on the id_en field, aligning the source and target texts for coherent translation.

To address the translation task, we employed the EasyNMT library, utilizing the 'opus-mt' model, renowned for its capability in handling complex linguistic structures. The model translated English puns into French while preserving their humorous essence, demonstrating its proficiency in maintaining the nuances of humor across languages.

Translations were performed on a subset of the training data, with the results compiled and saved in JSON format. These translated texts provided a clear indication of the model's capability to handle humor in translation, paving the way for further evaluation and analysis.

The translation of puns from English to French employed the EasyNMT model, which demonstrated a robust ability to maintain the humor and wordplay in translations.

3.4. Evaluation Methodology

We used BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) scores to evaluate the translation quality. These metrics are standard in machine translation for assessing how well the translations align with reference translations.

- BLEU Score: 6.35e-232 (This extremely low score suggests that the translations are significantly different from the reference translations. It indicates potential issues with translation accuracy and quality.)
- METEOR Score: 0.032 (This low score indicates that the translations did not align well with the reference translations in terms of content and structure.)

We conducted batch translations of training data to generate French translations for multiple English puns efficiently. This approach ensured that the model was exposed to a diverse set of humorous content, enhancing its ability to translate puns accurately. During the evaluation phase, we processed test queries systematically, iterating over each query to ensure precise translations of relevant puns.

The results indicate that while the EasyNMT model performed reasonably well in terms of human evaluation, the objective metrics (BLEU and METEOR scores) suggest significant challenges in translation quality. The low BLEU score, in particular, highlights issues with the model's alignment with reference translations. Future work will focus on refining the model and improving translation quality to better handle the complexities of humor in translation.

4 Conclusion

The findings from our study underscore the potential of a comprehensive approach in addressing the challenges of humor-aware information retrieval, classification, and translation. By integrating TF-IDF vectorization, Logistic Regression, and advanced neural translation models, we demonstrated the effectiveness of these techniques in processing and understanding humorous content.

Our results indicate that machine learning models can capture the nuances of humor to a significant extent. However, there are still areas that require further exploration and improvement. Future work should focus on enhancing the interpretability of models, addressing cultural and contextual dependencies in humor, and exploring more sophisticated techniques for handling the subtleties of humorous content.

This study contributes to the broader field of computational humor, offering insights into the potential and limitations of current machine learning models. By advancing our understanding of

humor analysis and translation, we pave the way for more effective and engaging human-computer interactions in the future.

Acknowledgments

We thank the CLEF 2024 organizers for their guidance and support. Special thanks to Liana Ermakova for her overview of modern methods and training students, including those with no coding background, as part of the BIP course AI for Humanitarians. We also thank the course organizers for providing the opportunity to exchange experiences and knowledge with students from various European universities.

References

[1]. L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, and A. Jatowt. 2024. Overview of JOKER - CLEF-2024 track on Automatic Humor Analysis. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Laure Soulier, Giorgio Maria Di Nunzio, Petra Galuščáková, Alba García Seco de Herrera, Guglielmo Faggioli, Nicola Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)

[2]. L. Ermakova, A.-G. Bosser, T. Miller, T. Thomas-Young, V. M. Palma Preciado, G. Sidorov, and A. Jatowt. CLEF 2024 JOKER lab: Automatic humor analysis. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, Proceedings, Part VI, volume 14613 of Lecture Notes in Computer Science (ISSN 0302-9743), pages 36–43, Cham, March 2024. Springer. ISBN 978-3-031-56072-9. DOI: 10.1007/978-3-031-56072-9_5.

[3]. L. Ermakova et al. 2024. Overview of the CLEF 2024 JOKER Task 1: Humor-aware information retrieval. In: Guglielmo Faggioli et al. (Eds). 2024. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org.

[4]. V. M. Palma Preciado et al. 2024. Overview of the CLEF 2024 JOKER Task 2: Humor classification according to genre and technique. In: Guglielmo Faggioli et al. (Eds). 2024. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org.

[5]. L. Ermakova et al. 2024. Overview of the CLEF 2024 JOKER Task 3: Translate puns from English to French. In: Guglielmo Faggioli et al. (Eds). 2024. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings, CEUR-WS.org.

[6]. L. Ermakova, A.-G. Bosser, A. Jatowt, and T. Miller. 2023. The JOKER Corpus: English-French Parallel Data for Multilingual Wordplay Recognition. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 2796–2806. https://doi.org/10.1145/3539618.3591885

[7]. J.D. Kelleher, B. Mc. Namee, and Aoife D'Arcy. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. Illustrated edition. The MIT Press, 2015. ISBN 0262029448, 9780262029445.

[8]. C. Sebastian. Python Programming for Beginners: Learn Python Machine Learning Language from Scratch, Deep Learning with Python. Amazon Digital Services LLC - KDP Print US, 2018. ISBN 1792874650, 9781792874659.

[9]. V.S.S. Chandra and A. Hareendran S. Artificial Intelligence and Machine Learning. PHI Learning, 2014. ISBN 8120349342, 9788120349346.