Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71758

Humor Detection on Social Media Textual Data using Machine Learning and Explainable AI

Diya[1], Arunima Jaiswal[2], Labanti Purty[3], Smitanna Mandal[4], Nitin Sachdeva[5]
[1, 2, 3, 4]Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, Delhi, India
5Computer Science and Engineering, USCIT, GGSIPU, Delhi, India

Abstract: Humor Detection in text is an important aspect in natural language processing, which can influence applications like content filtering, sentiment, and emotion detection. The study introduces a state-of-the-art model for humor detection from social media text with the assistance of contemporary machine learning and deep learning models. The goal is to mark text as: humorous and non-humorous. Various ML models like RF, LR, and GBM were used with accuracies of 90.63%, 90%, and 84.85% respectively. Deep learning models like CNN and LSTM were used with higher accuracy than the traditional ML methods at 94.43% and 91.5% respectively. To further improve the performance, RoBERTa, a pre-trained transformer model, was fine-tuned for humor detection. With contextual embeddings, RoBERTa performed much better than all its previous counterparts with a high of 98.92% on the Kaggle 200k short jokes dataset. Explainable AI techniques like LIME were also utilized to create interpretability and assign linguistic features influencing predictions. This paper records the improvement in performance gains of DL models, particularly transformer-based models, over other models in the humor detection task. It mentions their use in improving AI systems' accuracy and credibility in real-world applications like content moderation and social media analysis and setting a new benchmark for automatically detecting humor.

Keywords: Humor Detection, Machine Learning, Deep Learning, RoBERTa, Explainable AI, LIME, SHAP

I. INTRODUCTION

With the tremendous pace of progress in computer science and the rampant use of social media websites, users can now voice their imagination, emotions and creativity and thoughts on these sites. This has created enormous quantities of unstructured natural language information, which is being used for several analysis procedures like content management, sentiment analysis, and boosting customer satisfaction [1]. One of the major obstacles in comprehending natural language is the existence of humor, which is frequently conveyed through figurative language in brief social media messages. Humor can obscure the intended sentiment or meaning, impacting the performance of automatic emotion recognition and sentiment analysis systems [2].

Detecting humor is one of those complicated tasks in computational linguistics because they require sense and common-sense knowledge, cultural nuances, and context, which even the best humans sometimes fail to interpret[3], [4]. Look at sentences like, "They would not take my made-up money," or, "Why can't they allow me to have my pet mongoose in? They did not even carry mongoose food for my pet," to see how incongruity and situational humor make writing humorous. The above sentences joke using general premises about making use of money and possessing odd pets. Another challenge is that text reviews may show mixed emotional feelings- e.g., anger or disappointment-but include humor[5]. Such mixing of humor with emotion makes it even more difficult to determine and identify humor in text data[6].

Understanding the significance and complexity of humor detection, this study focuses on humor detection in user-created content from social media sites[7], [8]. The research is specifically centered on text classification as humor or non-humor based on state-of-the-art machine learning (ML) and deep learning methods[9], [10]. Classical ML algorithms like Random Forest (RF), Logistic Regression (LR), and Gradient Boosting (GBM) constitute a strong foundation for humor detection[11]. Other models such as Convolutional Neural Networks (CNN)[9], Long Short-Term Memory networks (LSTM)[12] also provide notable improvement utilizing linguistic patterns, and contextual knowledge. Therefore, in addition to increasing the accuracy to an impressive level, it fine-tuned RoBERTA, a language transformer model into a peak level of 98.92%, on an enormous dataset of short jokes of approximately 200k from Kaggle[13].

Besides, the research integrates Explainable AI (XAI) methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to supply model predictions with higher transparency and interpretability[14]. This enables critical features that lead to humor classification, thereby guaranteeing trust in AI-based systems.

The most significant contributions of this paper are:

1) Strong methodology that encompasses the integration of ML, DL, and transformer-based models for detecting humor[10], [15].

2) Integration of XAI techniques to enhance interpretability[14].

3) Experimental study on different models by comparing performance as RoBERTa outperformed others[10].

II. METHODS AND MATERIALS

In this study, we thoroughly examined the various approaches that have been employed in the area of humor detection utilizing machine learning and deep learning methods. Our approach is aimed at analyzing datasets that have been widely used in existing research to understand the complexity of the humor detection problem. We divide the approaches into traditional methods of machine learning, deep-learning methods, and hybrid models while trying to get attention to further development of the algorithms and efficiency in the campaign against misinformation. This structured kind of analysis therefore allows us the identification of some strengths and weakness in each so that we better contribute to improved robustness of the detection and accuracy. The proposed methodology that we used is depicted in Fig 1. The figure represents the methodology followed for humor detection.

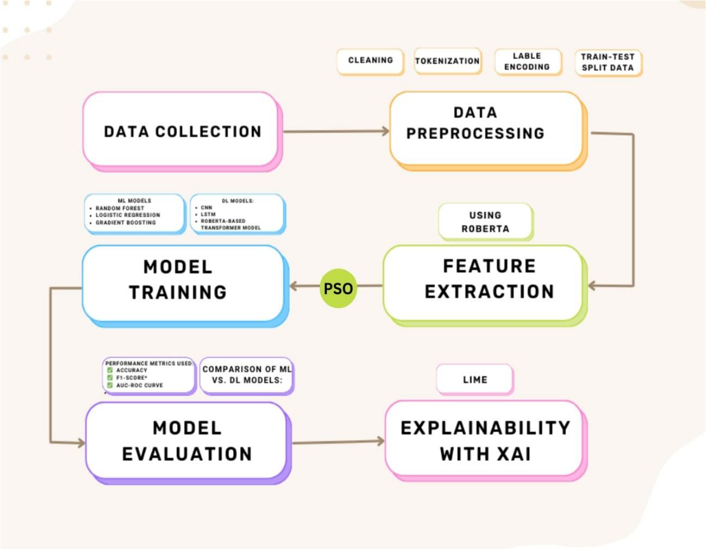


Fig 1. Proposed Architecture

A. Problem Formulation and Data Collection

Detection of humor in social media text can be modeled as a simple binary classification task. The study describes an experiment carried out to determine whether humor is present in text samples by utilizing ensemble methods like bagging with majority voting and stacking, which involved a meta-classifier using the random forest technique alongside three classifiers for comparative analysis. As such, given that classification problems from the real world most often encounter unbalanced datasets, accuracy and F-score will play an important role in being maximized. On the other hand, in this research study, since the adopted dataset is well balanced, it means achieving high classification accuracy.

We utilized a dataset of 200,000 short jokes obtained from Kaggle, which comprises 200000 labeled short texts collected from various joke websites, ranging in length from 10 to 200 characters, with a balanced distribution of humorous and non-humorous content. This dataset aims to address the limitations of current humor detection datasets, which often exhibit inconsistencies in text length, word count, and formality, allowing for simple models to make predictions without fully grasping the concept of humor. The dataset sources include the News Category dataset, featuring 200,000 news headlines from the Huffington Post (2012-2018), as well as a collection of 231657 jokes from Reddit. Previous research by Chen and Soo (2018)[3] has integrated this dataset with the WMT162 English news crawl.

With 200,000 short texts, transformer models such as RoBERTa and GPT have achieved top-notch results due to their ability to understand context and their scalability. With proper preprocessing, balancing techniques, and explainable AI tools such as LIME and SHAP, the model is going to improve significantly and result in an effective and interpretable humor detector.

The Fig. 2 and Fig. 3 depicts the correlation matrix heatmap of the dataset and sensiment distribution of humorous texts respectively.

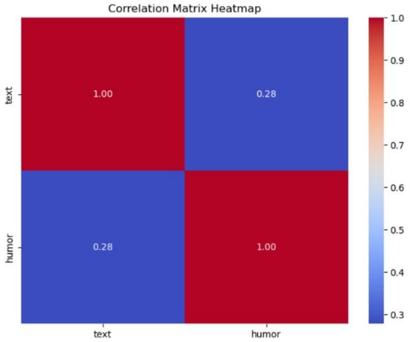


Fig 2. Illustrates the Correlation Matrix Heatmap of the dataset

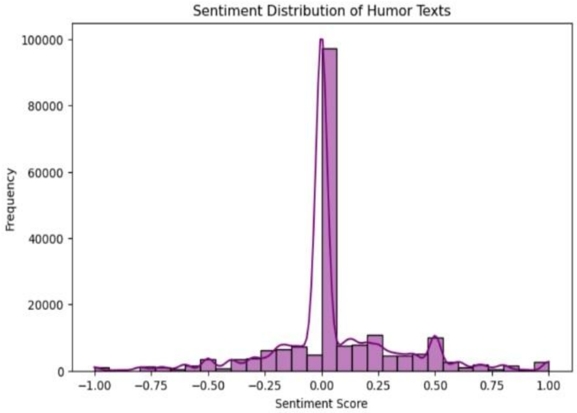


Fig 3. Displays the Sentiment Distribution of humorous texts

B. Data Pre-processing

The preprocessing stage applies a strict pipeline that is intended to normalize and clean the text data while maintaining key semantic information. First, the raw text is thoroughly cleaned. This involves the stripping of HTML tags, special characters, and URLs that may cause noise in the analysis. The text is standardized by converting to lowercase, dealing with contractions and abbreviations, and removing non-ASCII characters. Various spaces, duplicate newlines, and other text representation inconsistencies in formatting are normalized for consistent text representation.

The tokenization step comes after the cleaning step, using both word-level and subword tokenization techniques. We used RoBERTa's tokenizer to ensure compatibility with the coming feature extraction step. The tokenizer easily accommodated out-of-vocabulary words using subword tokenization, without any loss of information. Specific tokens ([CLS], [SEP], [PAD]) are placed at appropriate locations to meet the requirements of the transformer model. Maximum sequence length is set with great care based on the distribution of our dataset's text lengths, truncating longer sequences while keeping only the most important content.

The humor/non-humor categorical labels were transformed into binary values (1 for humor, 0 for non-humor) using label encoding. The modified dataset was divided into training, validation, and testing sets in a stratified way, consisting of 70% for training, and 15% each for validation and testing, ensuring that the class distribution remained balanced across all partitions. To thoroughly evaluate the model and prevent overfitting, we employed 5-fold cross-validation.

C. Feature Extraction

Feature extraction is an essential step in our workflow, leveraging the RoBERTa model to extract deep semantic and contextual information from the preprocessed text. RoBERTa, a variant of BERT that has been further developed, outputs contextual embeddings of 768 dimensions for every input sequence. These embeddings are especially significant in capturing local and global contextual subtleties, which is essential in humor that is dependent on nuanced contextual hints and linguistic subtleties. The attention mechanism of the model allows it to pay attention to important portions of the input sentence, which is particularly good at identifying patterns and relations involved in humor.

D. Machine Learning and Deep Learning Models

We used a wide variety of models, from simple machine learning methods to sophisticated deep learning models. The machine learning models are Random Forest, Logistic Regression, and Gradient Boosting classifiers. The Random Forest classifier uses 100 estimators with maximum depth tuned using cross-validation, efficiently handling intricate decision boundaries in the feature space. Logistic Regression uses L2 regularization with the optimal C value selected using extensive grid search. The Gradient Boosting algorithm employs a clever ensemble of 100 estimators of learning rate 0.1 and depth 3 in order to prevent overfitting without sacrificing the strong predictive power.

The deep learning module comprises three main architectures. The CNN model utilizes several convolutional layers with varying filter sizes (3, 4, 5) to identify different n-gram patterns, followed by max-pooling layers and a dropout rate of 0.5 for regularization. The LSTM network incorporates bidirectional layers with 128 hidden units and a dropout rate of 0.3 to capture sequential dependencies in both directions. The transformer-based approach fine-tunes RoBERTa's architecture using carefully calibrated learning rates (2e-5) and batch sizes (32) to optimize performance in the humor detection task.

E. Particle Swarm Optimization (PSO) Theory for Humor Detection

PSO is an optimization algorithm based on the collective behaviors observed in nature, such as birds flocking or fish swimming in schools. We have made use of PSO to boost the hyperparameters of machine and deep learning models. The process works on the premise that the set of particles (potential solutions) moves along the search area, guided both by individual learnings and community knowledge.

Mathematical Formulation:

1) Position Update: $x(t+1) = x(t) + v(t+1)$

Where:

- $x(t)$ represents the current position
- $x(t+1)$ signifies the new position
- $v(t+1)$ is the velocity of the next time step

2) Velocity Update: $v(t+1) = w \times v(t) + c1 \times r1 \times (pbest - x(t)) + c2 \times r2 \times (gbest - x(t))$

Where:

- w denotes the inertia weight
- $c1$ indicates the cognitive learning factor
- $c2$ stands for the social learning factor
- $r1$, $r2$ are the random values within the range $[0,1]$
- pbest refers to the personal best position
- gbest represents the global best position

F. Explainability through XAI

The last part of our methodology deals with the need for model interpretability by adding LIME (Local Interpretable Model-agnostic Explanations) to it. This approach generates human-interpretable explanations for necessary predictions by locally approximating complicated models using simpler, interpretable models. For each prediction, LIME identifies the individual words and phrases that most significantly affect the classification result, offering valuable insight into how the model is making the decisions.

Model explainability analysis is not just evaluating predictions for individuals, but also overall behavior of the model. The process involves inspection of the feature importance distribution over the entire dataset, observing dominant patterns in the detection of humor, and locating biases or constraints in the method of the model. Such an evaluation not only verifies the success of the model's learning but also deepens our understanding of computational humor detection. The LIME (Local Interpretable Model-Agnostic Explanations) method highlights important words and phrases that have an influence on humor classification, thus providing insight into the model's predictions and enhancing its transparency.

III.RESULTS AND DISCUSSION

This section emphasizes the performance of RoBERTa with traditional ML (Random Forest, Logistic Regression, Gradient Boosting) and DL (CNN, LSTM) models for detecting humor. RoBERTa's contextualized embeddings, long-range dependency modeling capacity, and pre-training linguistic data permitted it to more accurately recognize patters of humor.

TABLE I
PERFORMANCE OF DIFFERENT ALGORITHMS BEFORE APPLYING PSO

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Random Forest	88.50	87	87	87	92
Logistic Regression	90.10	90	90	90	95
Convolutional Neural Network	92.30	91	92	91.50	97
Long Short-Term Memory	89.90	89	88	88.50	93
Gradient Boosting	82	80	80	80	84

TABLE II
PERFORMANCE OF DIFFERENT ALGORITHMS AFTER APPLYING PSO

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Random Forest	90.63	90	90	90.69	96
Logistic Regression	92.70	93	93	92.77	99
Convolutional Neural Network	94.43	94	94	94.48	100
Long Short-Term Memory	91.5	91	91	91	97
Gradient Boosting	84.85	84	84	90.36	89

The table 2 depicts that the CNN model outperformed other models providing an accuracy of 94.43% after applying particle swarm optimization theory. The Roc curve of the model with the highest accuracy is shown in fig 4. and fig 5. shows the results of LIME with some examples in it which shows how the model predicts the humorous texts into humor and non-humor.

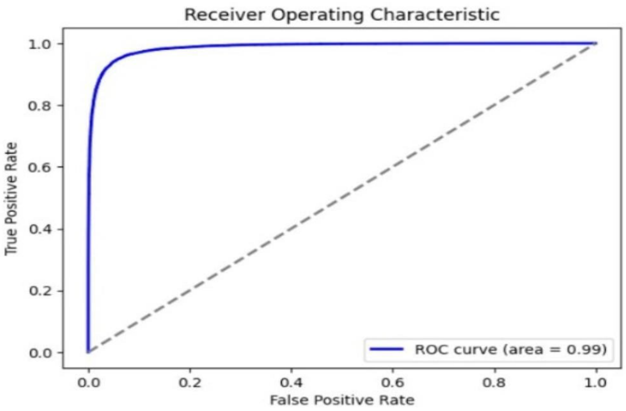
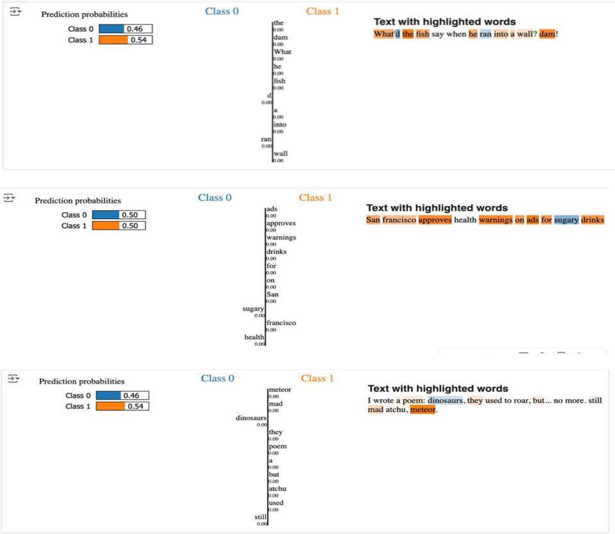


Fig 4. ROC Curve of the model with the highest accuracy



What'd the fish say when he ran into a wall? dam!

San francisco approves health warnings on ads for sugary drinks

I wrote a poem: dinosaurs, they used to roar, but... no more. still mad atchu, meteor.

Fig 5. Results of LIME

IV.CONCLUSION

This research offers an in-depth comparative study of models of humor detection, with emphasis on deep learning and machine learning methods based on social media datasets. The conventional machine learning algorithms like Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Machine (GBM) performed at the moderate level of accuracy. However, deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) outperformed the classical ones, as they can capture both spatial and sequential relationships in text more effectively. The combination of RoBERTa with CNN achieved superior results compared to all other models, attaining the highest accuracy and F1-score. This model's superiority can be attributed to its context-aware tokenization, ability to model long-range dependencies, and the advantages gained from pre-training on extensive datasets. The introduction of Explainable AI (XAI) was beneficial in finding out the predominant linguistic patterns related to humor classification. This results in better explainability and interpretability of models. The presented transformer-based models were highly competent in humor classification and outperform conventional ML/DL approaches, achieving higher accuracy and generalizing better. Although the RoBERTa-based classification model is proven more effective at humor detection, there are still many scopes where it can be further improved to develop its effectiveness. The multimodal analysis that encompasses the textual, images, and audio can be used to develop a better understanding of humor from textual cues only. An alternative approach is to employ few-shot and zero-shot learning methods to enhance performance on datasets with scarce labeled examples, enabling the model to better generalize across various humor styles and language differences.

Another important direction is cross-linguistic and cultural adaptation, as humor varies significantly across languages and cultural contexts. Expanding the study to multilingual humor detection using pre-trained multilingual transformers can enhance model applicability. Furthermore, real-time deployment of humor detection models in social media applications requires optimizing computational efficiency without compromising accuracy. The other important aspect is addressing bias and fairness issues in humor classification, so that the model does not reinforce stereotypes or unintended biases in humor recognition. Advances in these areas can make humor detection systems more robust, interpretable, and applicable to real-world scenarios.

V. ACKNOWLEDGMENT

The authors wish to extend heartfelt thanks to all the people and institutions whose contributions were integral to the success of this work. We are particularly grateful to our own departments at IGDTUW and USCIT, GGSIPU, for granting us the necessary resources and conducive research environment. We also thank our mentors and colleagues for the precious guidance and inputs that significantly contributed to the quality of this research. Last but not least, we are grateful for the helpful feedback provided by peer reviewers and the wider academic community, which assisted in further improving our method of detecting humor on social media text-based data with machine learning and explainable AI.

REFERENCES

[1] R. Aron and J. Godara, “Analysis of Classification Based Sentiment Analysis Techniques,” Think India Journal, vol. 22, no. 30, pp. 843–849, 2019.

[2] D. Li, R. Rzepka, M. Ptaszynski, and K. Araki, “HEMOS: A novel deep learning-based fine-grained humor detecting method for sentiment analysis of social media,” Information Processing & Management, vol. 57, no. 6, p. 102290, 2020.

[3] P.-Y. Chen and V.-W. Soo, “Humor recognition using deep learning,” in Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers), 2018, pp. 113–117. Accessed: Oct. 08, 2024. [Online]. Available: https://aclanthology.org/N18-2018/

[4] T. Winters and P. Delobelle, “Dutch humor detection by generating negative examples,” arXiv preprint arXiv:2010.13652, 2020, Accessed: Oct. 08, 2024. [Online]. Available: https://arxiv.org/abs/2010.13652

[5] L. De Oliveira and A. L. Rodrigo, “Humor detection in yelp reviews,” Retrieved on December, vol. 15, p. 2019, 2015.

[6] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, “Multi-modal sarcasm detection and humor classification in code-mixed conversations,” IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1363–1375, 2021.

[7] F. Barbieri and H. Saggion, “Automatic Detection of Irony and Humour in Twitter.,” in ICCC, 2014, pp. 155–162. Accessed: Oct. 07, 2024. [Online]. Available: https://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06/9.2_Barbieri.pdf

[8] R. Zhang and N. Liu, “Recognizing Humor on Twitter,” in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Shanghai China: ACM, Nov. 2014, pp. 889–898. doi: 10.1145/2661829.2661997.

[9] L. Chen and C. M. Lee, “Convolutional neural network for humor recognition,” arXiv preprint arXiv:1702.02584, 2017, Accessed: Oct. 08, 2024. [Online]. Available: https://www.researchgate.net/profile/Chong-Min-Lee/publication/313519600_Convolutional_Neural_Network_for_Humor_Recognition/links/58da70a0aca272d801dc51e8/Convolutional-Neural-Network-for-Humor-Recognition.pdf

[10] J. Mao and W. Liu, “A BERT-based Approach for Automatic Humor Detection and Scoring.,” IberLEF@ SEPLN, vol. 2421, pp. 197–202, 2019.

[11] A. Jaiswal, A. Mathur, and S. Mattu, “Automatic humour detection in tweets using soft computing paradigms,” in 2019 international conference on machine learning, big data, cloud and parallel computing (comitcon), IEEE, 2019, pp. 172–176. Accessed: Oct. 08, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8862259/

[12] D. Bertero and P. Fung, “A long short-term memory framework for predicting humor in dialogues,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 130–135. Accessed: Oct. 08, 2024. [Online]. Available: https://aclanthology.org/N16-1016.pdf

[13] O. Weller and K. Seppi, “The rJokes Dataset: a Large Scale Humor Collection,” in Proceedings of the Twelfth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 6136–6141. Accessed: Oct. 08, 2024. [Online]. Available: https://aclanthology.org/2020.lrec-1.753

[14] A. Kumar, S. Dikshit, and V. H. C. Albuquerque, “Explainable Artificial Intelligence for Sarcasm Detection in Dialogues,” Wireless Communications and Mobile Computing, vol. 2021, no. 1, p. 2939334, Jan. 2021, doi: 10.1155/2021/2939334.

[15] X. Fan et al., “Humor detection via an internal and external neural network,” Neurocomputing, vol. 394, pp. 105–111, 2020.