THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Humor Analysis on Online Crowdsourced Content

Avery Pratt
Spelman College

Malcolm Felix
Prairie View A&M University

Scarlett(Jiaqi) He
University of Chicago

Casidhe Pierre
Florida A&M University

Dickson Acheampong
Howard University

Abstract

This project explores how humor in online content can perpetuate bias and discrimination. To address this, we've developed a novel annotation framework that labels jokes across three dimensions: classification type, rhetorical device, and targeted group. Using both instruction-tuning and fine-tuning techniques, we trained models to automate this labeling. The instruction tuning team experimented with large language models like Mistral and OpenAI to classify jokes using few-shot prompting. Meanwhile, the fine-tuning team applied transformer-based models such as DistilBERT and DeBERTa to improve performance on specific label types. Our study highlights the challenges of annotating humor and proposes scalable solutions to improve automated detection of harmful jokes.

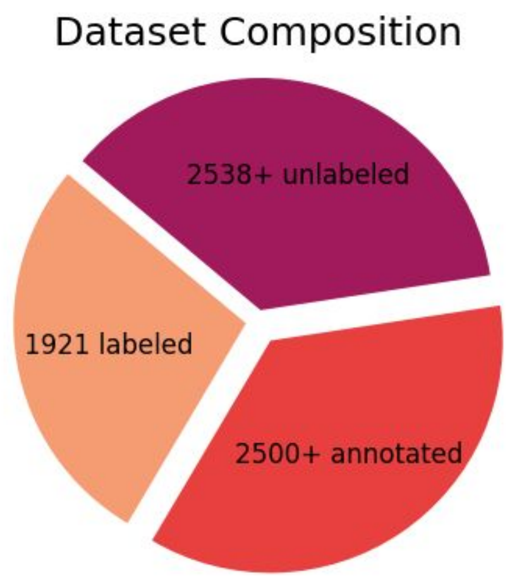
Problem

Humor is a powerful tool in online communication, but it often conceals bias, stereotypes, and harmful messaging. Detecting these patterns is challenging because humor is subjective and context-dependent. To address this our work makes the following contributions:

- A novel annotation framework that categorizes humor into fine-grained labels
- Fine-tuned models to test the feasibility of the proposed annotation framework in practice

Data Overview

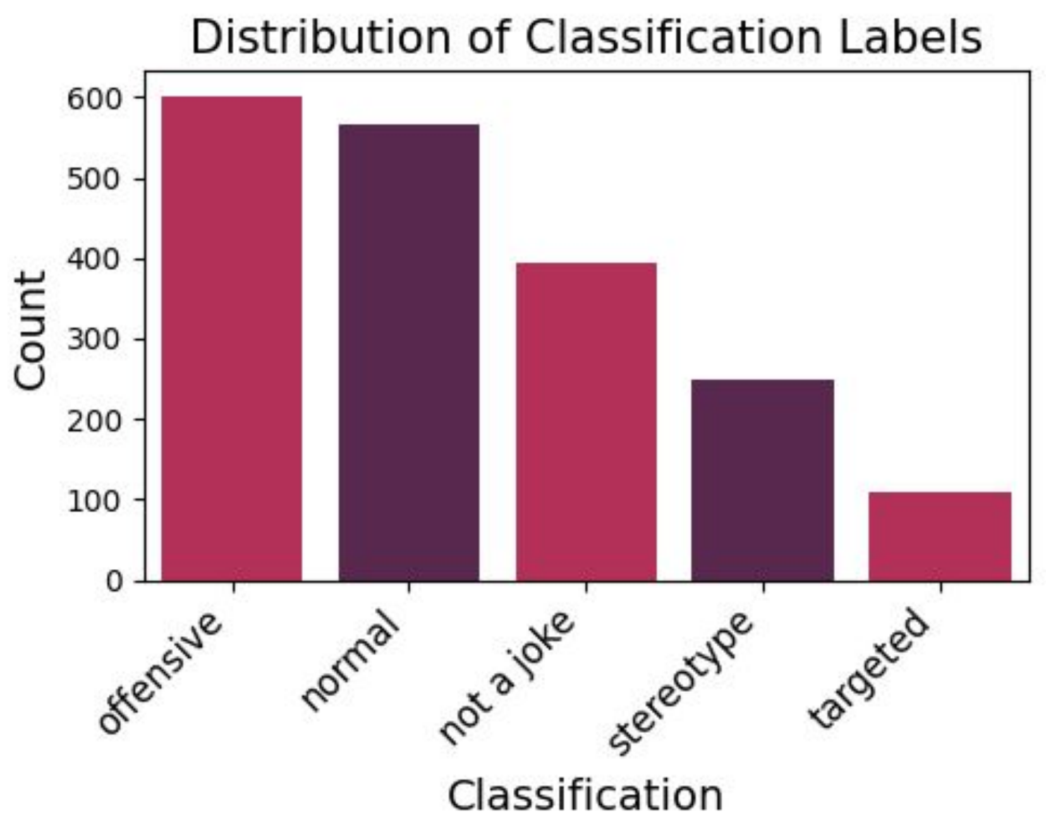
- The dataset includes approximately 3,500 jokes sampled from a dataset of 100,000+ jokes from Kaggle
  - 1,921 labeled jokes for training
  - 2,538 unlabeled jokes for testing
- We produced additional labeled data of 2,500 jokes following the annotation framework proposed below

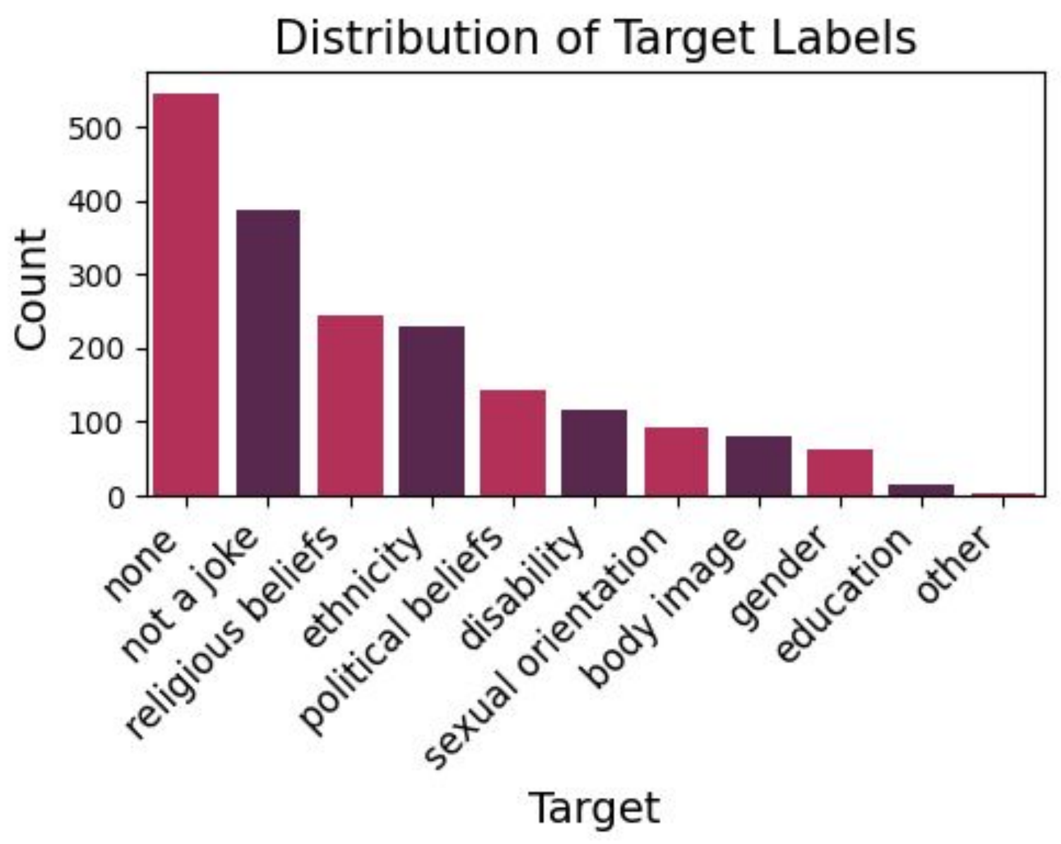


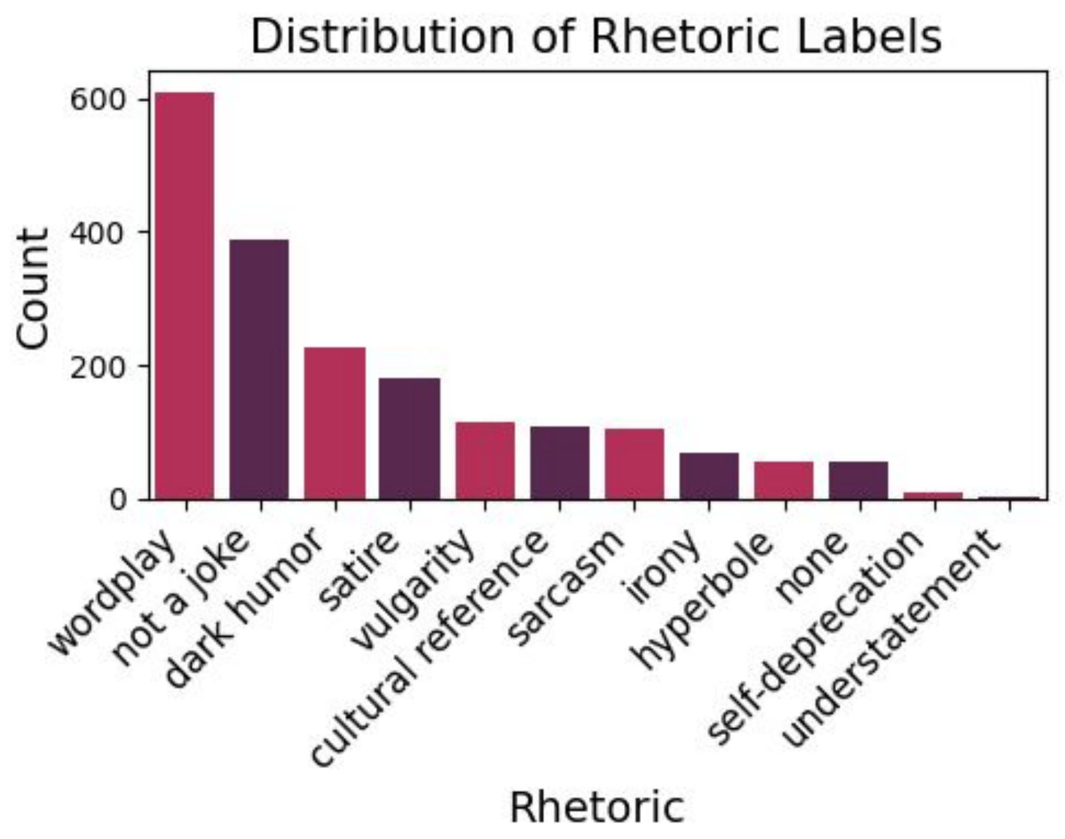
Categories	Classification Tasks
Classification Label - 5 Labels	Binary & MultiClass
Target - 11 Labels	MultiClass
Rhetoric - 15 Labels	MultiClass

Table 1, Annotation Framework

Annotation Framework







Classification Label: categorization over humorous and regular texts, with additional fine-grained labels for jokes

Target: considers whether the joke is directed towards any specific demographics

Rhetoric: specifies the linguistic technique used to deliver a joke

Instruction Tuning

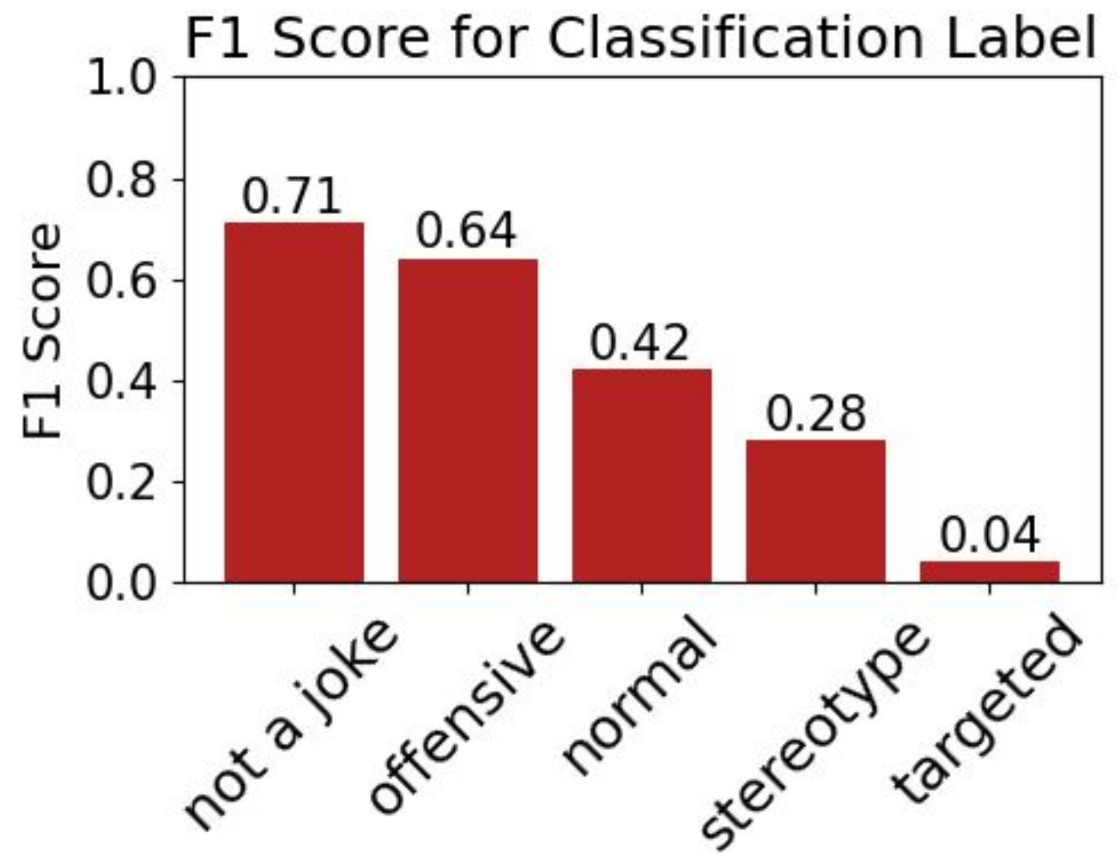
Methodology

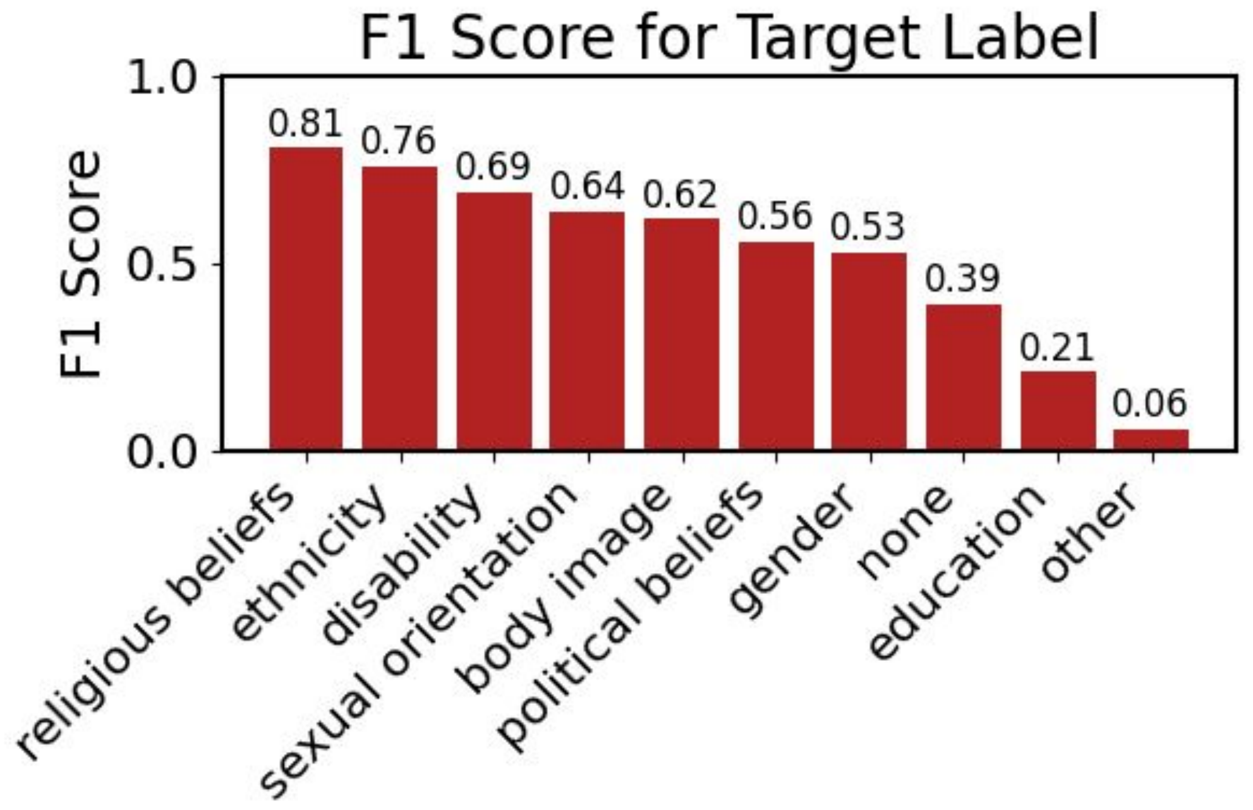
- Developed 3 separate humor classifiers one for each category
- Required structured prompts of task description,example text and label with examples processed in chat completion format
- At inference time, model took in task description and example text. Generated label as new tokens.
- Used few-shot prompting for Open AI and Mistral but larger labelled dataset for Gemma and Phi.

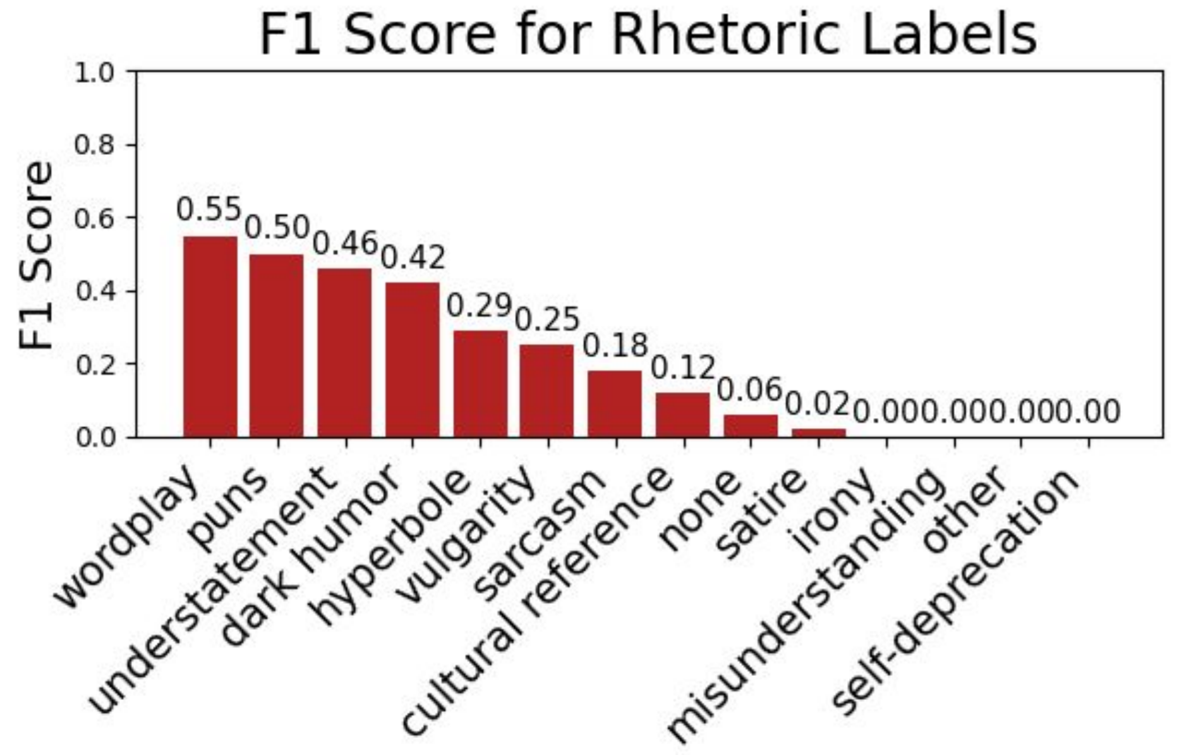
Models Tested:

1. Open AI
2. Mistral
3. Gemma
4. Microsoft Phi

OpenAI Results







Key Takeaways

- Split data into training and validation; evaluated using F1-score to account for class imbalance
- **OpenAI** performed well but struggled with jokes that had classification label *targeted* or target label *education*
- Models consistently found the linguistic nuances of the rhetoric category difficult to generalize

Fine Tuning

Methodology

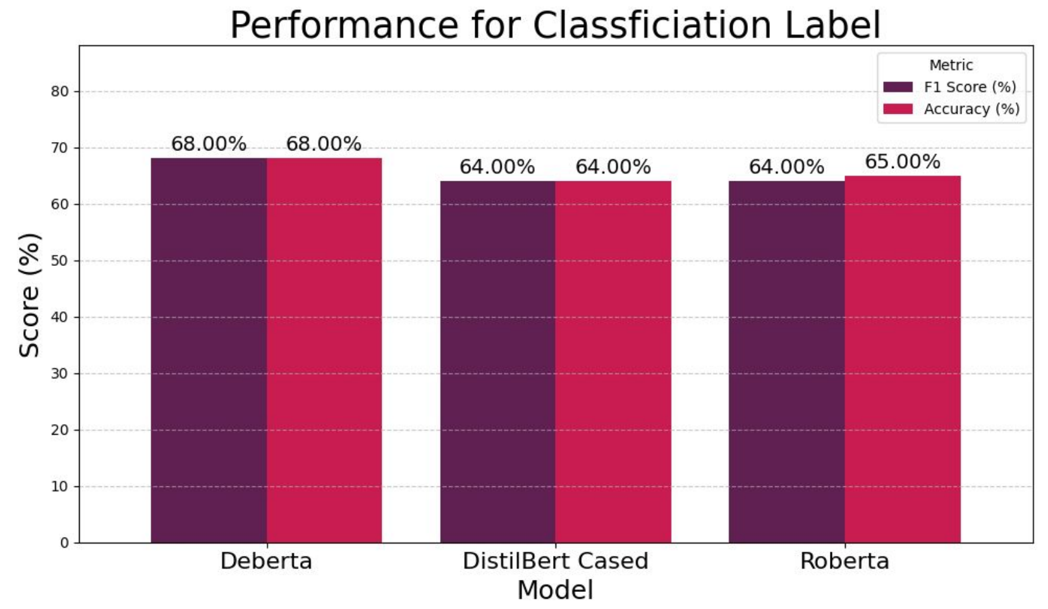
- Pre-trained models optimized for humor classification
- Hyperparameter Tuning (Batch Sizes, Learning Rates, etc.) and Class Weights Balancing to optimize performance

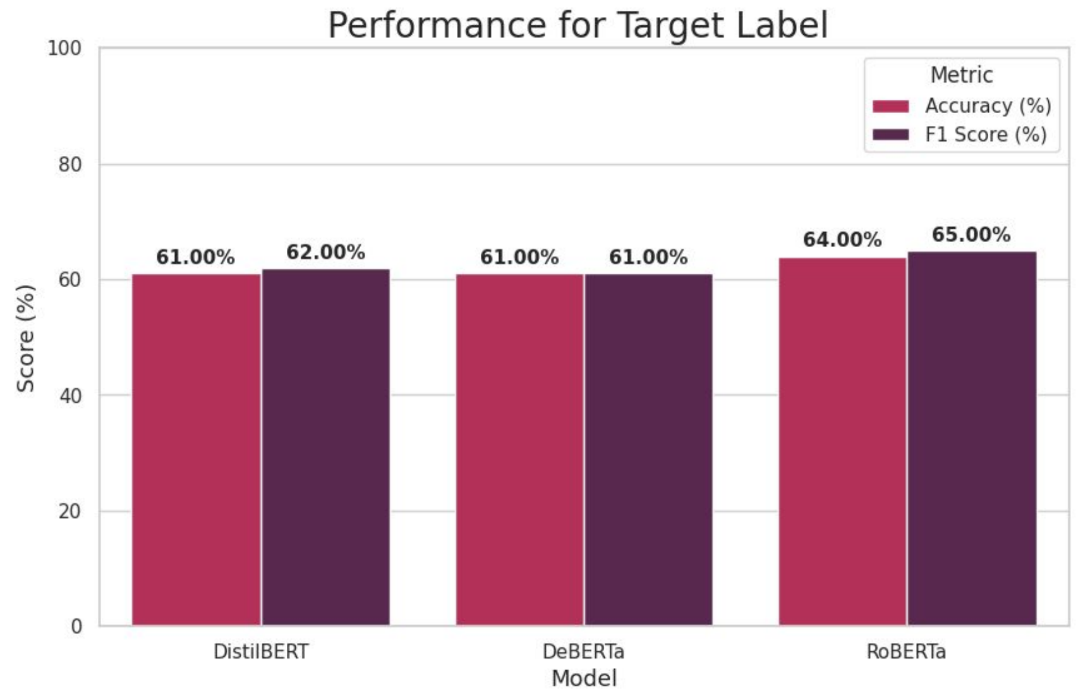
Models Tested:

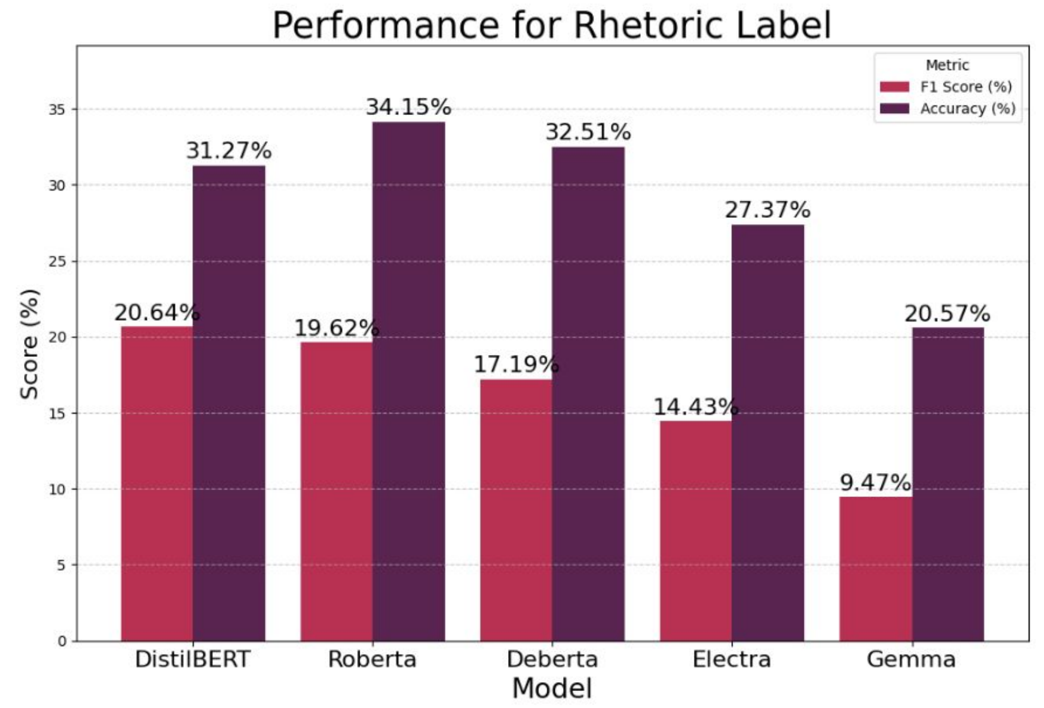
1. DistilBert
2. Roberta
3. Electra
4. Gemma
5. Deberta
6. Microsoft Phi

Key Takeaways

- **DeBERTa** performed the best on Classification Labels
- **RoBERTa** performed the best on Target
- **DistilBERT** performed the best on Rhetoric







Conclusions

Our study confirms that humor classification is challenging but feasible with the right structure and tools. Prompt engineering offers flexibility and speed, especially for prototyping or exploring new joke types. However, fine-tuned models generally perform better on fixed tasks with well-labeled datasets. Both approaches benefit from detailed annotation guidelines and diverse human input. By combining these strategies, we move closer to automated systems that can understand and mitigate bias hidden in online humor.

Acknowledgements

We would like to acknowledge Dr. Satadisha Saha Bhowmick, Dr. Patricia Chiril, Summer Han, Dr. Evelyn Campbell, & Data Science Institute at UChicago.