



Can jokes help people understand why a piece of news is fake?

Hugo Damião Ferrão Ribeiro

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisor: Prof^ª Helena Sofia Andrade Nunes Pereira Pinto

Examination Committee

Chairperson: Prof. Nuno Miguel Carvalho dos Santos
Supervisor: Prof^ª Helena Sofia Andrade Nunes Pereira Pinto
Member of the Committee: Prof. Nuno João Neves Mamede

October 2024

Acknowledgments

I would like to thank my family, my parents and grandparents for their encouragement and support over all these years, for always being there for me through thick and thin and without whom I would not be where I am today.

I would especially like to thank my girlfriend for being my emotional support human. She's always there to listen to my complaints, and I know I can count on her to challenge my patience in the best way possible! Without a doubt, she has been the best surprise of this entire journey.

I would also like to thank my dissertation supervisor Prof. H. Sofia Pinto for her guidance, support, sharing of knowledge and learning opportunities that have made this Thesis possible.

My gratitude also goes to my colleagues and friends. A special thanks to "Os Velhos Sábios", who made my university experience much better with all the fun times we shared. I'd also like to thank Martim and João for their support. From game nights to "dailies" and the procrastination, even when we should have been working, these are moments I will forever cherish. The same is true for all my "Bruvah's" with whom I played games so many times on Discord, helping me relax after long days of work.

To each and every one of you – Ty.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Abstract

In an era marked by the rapid dissemination of information through various media channels, the challenge of discerning between credible news and misinformation is increasing significantly. To fight this, the CIMPLE project was created to make AI-powered explanations of misinformation. This thesis, part of this project, explores an approach to enhancing news understanding by integrating humor into explanations. The research aims to develop a system that transforms news explanations into humorous sentences, making them more engaging and memorable. Through this, we hope the heightened engagement and improved information retention will help individuals better discern and resist misinformation. To achieve this, we conducted a comparative study to identify the most suitable model from a range of options, which we then used to build our system. This system processes a claim, its label, and a human-crafted explanation, employing three distinct approaches for joke generation to produce an output that is humorous and memorable while preserving the explanatory aspect. Additionally, we developed and incorporated a moral and joke classifier to enhance the system's performance. We evaluated the system through a survey designed to measure how humorous explanations affect user perception and comprehension. Results showed that jokes generated by our system were often perceived as human-written, funnier, and more suitable than human-crafted jokes for explaining misinformation. Moreover, the findings indicated that creative content like humor was especially effective in explaining complex information, making people more receptive and open to engaging with it.

Keywords

Misinformation, Fake News, Jokes, Creative Explanations

Resumo

Num período marcado pela rápida disseminação de informação através de vários canais de comunicação, o desafio de distinguir entre notícias credíveis e desinformação aumenta significativamente. Para combater este desafio, o projeto CIMPLE foi criado com o objetivo de fornecer explicações de desinformação geradas artificialmente. Esta tese, inserida neste projeto, explora uma abordagem para melhorar a compreensão das notícias, integrando humor nas explicações. A investigação visa desenvolver um sistema que transforma explicações de notícias em frases humorísticas, tornando essas explicações mais apelativas e memoráveis. Para alcançar este objetivo, realizámos um estudo comparativo para identificar o modelo mais adequado entre várias opções, que posteriormente utilizámos para construir o nosso sistema. Este sistema processa uma notícia, a sua veracidade e uma explicação elaborada por humanos, utilizando três abordagens distintas de geração de piadas para produzir um resultado que seja simultaneamente humorístico e memorável, sem comprometer o aspeto explicativo. Além disso, desenvolvemos e incorporámos um classificador de moralidade e piadas para melhorar o desempenho do sistema. Desenvolvemos um questionário para medir o impacto das explicações humorísticas na perceção e compreensão dos utilizadores. Os resultados mostraram que as piadas geradas pelo nosso sistema eram frequentemente percebidas como escritas por humanos, mais engraçadas e mais adequadas do que as piadas criadas por humanos para explicar a desinformação. Além disso, os resultados indicaram que conteúdos criativos, como o humor, foram particularmente eficazes na explicação de informação complexa, tornando as pessoas mais recetivas e abertas a interagir com ela.

Palavras Chave

Desinformação, Notícias Falsas, Piadas, Explicações Criativas

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Work Objectives	3
1.3	Accomplishments	4
1.4	Document Structure	5
2	Background	6
2.1	Deep Learning Architectures	7
2.1.1	Recurrent Neural Networks	7
2.1.1.A	Long Short-Term Memory Networks	7
2.1.2	Sequence-to-Sequence Modeling	7
2.1.3	Attention Mechanism	8
2.1.4	Transformer	8
2.1.4.A	Prompt Engineering	9
2.1.4.B	Fine-Tuning	10
2.1.4.C	Hyperparameters	10
	A – Network Structure Hyperparameters:	10
	B – Training Algorithm Hyperparameters:	11
	C – Inference Process Hyperparameters:	11
2.2	Computational Creativity	11
2.3	Humor	12
2.3.1	Humor theories	12
2.3.2	Humor styles	13
2.3.3	Relation between Humor and Language	14
3	Related Work	15
3.1	Joke Generation Systems	16
3.1.1	Template-based Systems	16
3.1.2	Neural Systems	21

3.1.3	Hybrid Systems	25
3.2	Misinformation Fighting Systems	26
3.3	Humor Classification Systems	28
4	Problem Approach	32
4.1	Creative Explanations	33
4.2	Approach	33
4.2.1	Humor Generation Approaches Selection	34
4.2.2	System Components Overview	35
4.2.3	Result Analysis	35
5	Development	37
5.1	Comparison Study	38
5.1.1	Model List	38
5.1.2	Development and Results	39
5.1.2.A	GPT-2	39
5.1.2.B	GPT-Neo	40
5.1.2.C	T5	41
5.1.2.D	GPT-3	42
5.1.2.E	GPT-4	43
5.1.2.F	Conclusion	43
5.2	Joke generation Approaches	44
5.2.1	Prompt Chaining	44
5.2.2	First Fine-tuning Approach	46
5.2.2.A	Dataset	47
5.2.2.B	Fine-Tuning	48
5.2.2.C	Prompt Engenering	48
5.2.2.D	Hyperparameter Optimization	49
5.2.2.E	Domain-Specific Fine-Tuning	50
5.2.3	Second Fine-tuning Approach	50
5.2.3.A	Dataset	51
5.2.3.B	Fine-tuning	52
5.2.3.C	Hyperparameter Optimization	52
5.3	Moral Classifier	53
5.3.1	Moral Foundations	54
5.3.2	Dataset	54
5.3.3	Training and evaluation	55

5.3.4	Example of Morally Inappropriate Joke Detection	57
5.4	Joke Classifier	58
5.4.1	Classifier Options	58
5.4.1.A	Fine-tuned BERT Classifier	58
5.4.1.B	GPT-4 Powered Classifier	59
5.4.1.C	SSTH based Classifier	59
5.4.2	Performance Comparison	60
5.5	System Integration and Assembly	62
6	Questionnaire Evaluation	64
6.1	Structure	65
6.2	Results	68
6.2.1	Demographic and Humor preferences Data	68
6.2.2	Single Exercises	70
6.2.2.A	Human Written Jokes	70
6.2.2.B	Machine Generated Jokes	74
6.2.2.C	Effectiveness of Jokes in Fighting Misinformation	78
6.2.3	Comparison Exercises	83
6.3	Evaluation Discussion	89
7	Conclusion	90
7.1	Conclusions	91
7.2	Future Work	91
	Bibliography	91
A	Examples of Jokes Generated by our System	98
B	Survey Structure and Content Overview	102

List of Figures

3.1	Template-base system proposed by Petrovic et al.	16
3.2	Template-based system proposed by Yamane et al.	19
3.3	Types of joke production of the template-based system proposed by Sjobergh et al.	20
3.4	Template for the jokes of the system proposed by Labutov et al.	22
3.5	Architecture of a humor classifier based on a popular linguistic theory of humor.	29
4.1	Architecture of our joke generation system.	34
5.1	Evaluation of prompts and algorithm modifications.	45
5.2	Worcloud of our Dataset.	48
5.3	Percentage of Respondents Choosing Each Approach Results: Only one prompt (blue) vs. Two prompts (red).	49
5.4	Moral Classifier Metrics.	56
5.5	Moral Classifier Architecture.	57
6.1	Demographic Data.	69
6.2	Results distribution of “Do you consider this explanation to be a joke?” for human-written jokes.	70
6.3	Results distribution of “Do you consider this explanation to be a joke?” for human-written jokes.	71
6.4	Box chart illustrating the distribution of funniness ratings for human-crafted jokes.	71
6.5	Box chart illustrating the distribution of funniness ratings for human-crafted jokes.	72
6.6	Box chart illustrating the distribution of explainability ratings for human-written jokes.	73
6.7	Results distribution of “Do you consider this explanation to be a joke?” for machine- generated jokes.	75
6.8	Results distribution of “Do you consider this explanation to be a joke?” for machine- generated jokes.	75
6.9	Box chart illustrating the distribution of funniness ratings for machine generated jokes.	76

6.10	Box chart illustrating the distribution of funniness ratings for machine generated jokes. . .	76
6.11	Box chart illustrating the distribution of explainability ratings for machine generated jokes. . .	77
6.12	Results distribution of “How engaging do you find the joke compared to the explanation?”. . .	79
6.13	Results distribution of “Do you think you are more likely to remember the reason as to why the claim was classified with that Label?”.	79
6.14	Results distribution of “How likely would you be to read the full article after seeing the joke?”. . .	80
6.15	Results distribution of “How likely are you to share the joke?”.	81
6.16	Results distribution of “How likely are you to share the joke?”.	81
6.17	Results distribution of “How important is it for you to know the source of the information presented?”, (1 = “Not important at all”, 5 = “Extremely important”).	82
6.18	Results distribution of “After reading the joke, to what extent do you feel it tried to manip- ulate you or the information in any way?” (1 = “Not at all”, 5 = “Extremely”).	83
B.1	Introduction of the survey.	103
B.2	Humor Preference Questions.	105
B.3	First type of exercise (News article with machine or human-written joke).	106
B.4	Second type of exercise (News article without explanation and with a machine generated joke).	107
B.5	Comparison exercise between: Our system; Domain-Specific Fine-tuned Model and De- fault GPT-4.	108
B.6	Comparison exercise between: Machine and Human written joke.	109
B.7	Comparison exercise between: Jokes, Memes and Poems.	110

List of Tables

3.1	Joke generation systems reviewed.	31
5.1	Survey Results for GPT-2 and GPT-2 Large	39
5.2	Survey Results for GPT-Neo 1.7B and GPT-Neo 2.7B.	40
5.3	Survey Results for T5.	41
5.4	Survey Results for GPT-3Turbo and GPT-3 DaVinci.	42
5.5	Survey Results for GPT-4.	43
5.6	Number of jokes written by each person and their education degree.	52
5.7	Cohen Kappa Coefficients between Systems and Annotators.	61
B.1	Human Jokes used on our survey.	111
B.2	Machine Generated Jokes used on our survey.	113

Acronyms

CC	Computational Creativity
AI	Artificial Intelligence
DL	Deep Learning
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
LM	Language Model
LLM	Large Language Model
seq2seq	Sequence-to-Sequence
SSTH	Semantic Script Theory of Humor

1

Introduction

Contents

1.1	Motivation	2
1.2	Work Objectives	3
1.3	Accomplishments	4
1.4	Document Structure	5

The ease with which information can be created, disseminated, and accessed has reshaped our world. In the age of information, we find ourselves overwhelmed with a constant stream of news and claims. However, this very information landscape has given rise to a very concerning problem: Fake News and misinformation.

Most people have faced misinformation at some point in their lives, and in today's world, it often appears more appealing than the truth [1]. Additionally, explanations aiming to fight misinformation are typically tailored towards experts, resulting in content that can be perceived as excessively long, overly complex, filled with very detailed language, or even unengaging for the average person.

Axel Gelfert characterized fake news as “the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design.” [2]. The phrase “by design” is intended to emphasize that the misinformation is intentionally crafted, highlighting the intentional aim of fooling the audience for ulterior motives.

We can explore an example that illustrates the impact of fake news on public perception. One well known case is the 2016 Pizzagate incident, where baseless conspiracy theories circulated online, alleging a child trafficking ring linked to a Washington D.C. pizzeria. The rumors initially centered around false claims about Hillary Clinton being at the center of a pedophilia ring and they kept gaining momentum. As this was during the 2016 United States presidential election, rumors blew out of proportion. Shortly after the hashtag “#pizzagate” began trending on social media, it was used to spread false information, linking the alleged child trafficking ring to a Washington D.C. pizzeria, a place that John Podesta—Hillary Clinton’s campaign chairman—was known to frequent.

This led to a lot of harassment towards all the pizzeria’s employees and ultimately it escalated when a man walked in the pizzeria with a gun, motivated by the unfounded rumors, with the intent of saving the children being kept there [3]. Fortunately, no one was injured, but this incident shows the immense danger that fake news can pose to our society. It shows the need to fight the spread of misinformation in order to prevent more situations like this to happen.

1.1 Motivation

The CIMPLE project [4], dedicated to countering misinformation and fake news, aims to generate understandable explanations of Artificial Intelligence (AI) decisions and to use computational creativity to generate creative fake news explanations. This project explores a field within AI that focuses on enhancing the comprehensibility of AI systems for their users.

The field of explainability still holds immense growth potential. Taking this into consideration, we aim to forge a path in the right direction with our approach, which involves generating jokes using news

explanations as input. These news explanations, which play a pivotal role as the input for our system, are crafted by humans and sourced from dedicated misinformation-fighting websites with the purpose of illustrating how the information has been manipulated. It is fundamental to acknowledge that people process information in biased ways unknowingly [5]. Moreover, emotions play a notable role, changing people's attitudes toward science and influencing their risk management [6].

Humor, in this context, may be a useful tool in changing people's perspectives on specific topics. This happens as a consequence of making another person laugh. When a person reacts positively to a joke, they form a connection between the source of the humor and the pleasure of laughing. As a result of this link, the source of the humor is viewed as more likable [6].

Thus, we developed a system designed to rectify misinformation and counteract false narratives around a specific topic. We implemented a system that is trained to explain how information has been manipulated but in a humorous way. This unique capability allows the system to produce jokes as outputs, leveraging them as creative explanations that convey the rationale behind a given news article's label. In this system's framework, the input comprises a human-made news explanation, a claim, and its corresponding label (e.g., True, False, etc.). At the same time, the output becomes a humorous joke that enhances both its explanatory power and engaging appeal.

1.2 Work Objectives

Creative Explanations use imaginative techniques, making them more engaging and appealing to the audience. In our specific context, they entail the crafting of a joke, originating from a news explanation and its associated label, to deliver engaging insights. These explanations blend humor with more conventional informative elements [7]. These type of explanations are not easily created and present a unique challenge as they require an equilibrium between precise information and creativity.

With all this said, we state that our goals are:

- Using Computational Creativity (CC) be able to create explanations for Fake News;
- Develop a system that is capable of receiving an explanation, a claim and a label about a news article and use it to create a humorous text. The humorous part of the explanation has the objective of making it a more engaging experience for the readers;
- Create creative explanations that comply with the constraints set by the CIMPLE guidelines, in order to achieve good results. Our explanations, in the form of a joke, should be Informative, Creative, Accurate, and Appealing;
- Attaining a high resemblance with human-written jokes, where the generated results have the capability of passing as human-written, this means that when presented with two jokes, one human-

made and the other generated by a machine, a reader cannot easily tell which is which.

1.3 Accomplishments

Having established our objectives, it is essential to highlight the key accomplishments that reflect our progress toward these goals. These accomplishments illustrate the strides made in the pursuit of generating engaging, informative, and accurate explanations through humor.

Our main achievements can be outlined as follows:

- **Testing Various Models:** We tested a range of models to evaluate their humor generation capabilities, assessing factors such as joke relevance, coherence, and creativity. The tests revealed that certain models performed significantly better, providing a foundation for further fine-tuning;
- **Creation of a Dataset of Humorous Explanations for Fake News:** We compiled and curated a dataset that includes 100 humorous explanations designed to debunk fake news. This dataset serves as the basis for training and evaluating our models, ensuring they have diverse examples to learn from;
- **Successful Fine-Tuning of Machine Learning Models:** We fine-tuned multiple models, resulting in improved accuracy and humor quality. Fine-tuned models showed an improvement in producing jokes rated as “funny” by users, demonstrating the effectiveness of our adjustments;
- **Development of a Moral and Joke Classifier:** We developed a classifier capable of assessing the appropriateness of jokes and selecting the most suitable one from all generated options. This classifier ensures that the jokes remain respectful and optimally humorous, enhancing the system’s overall effectiveness and reliability;
- **Development of a System Capable of Generating Humorous Explanations:** We created a system that generates humorous explanations from a claim, a label, and an explanation;
- **Generation of Humorous Explanations Meeting Requirements:** The system consistently generated humorous explanations that met all imposed requirements, including creativity, funniness, and explainability;
- **Generation of Funnier Humorous Explanations than Human-Made Ones:** The system’s explanations were rated as funnier than those created by humans, with 6.7% more user evaluations considering the jokes generated by our system to be humorous. This indicates the system’s ability to surpass human performance in humor quality;

- **Generation of Humorous Explanations More Frequently Recognized as Explanations:** The system produced explanations that users identified as valid more often than human-made ones, achieving a recognition rate of 74.7%;
- **Generation of Humorous Explanations Providing Better Clarity:** Our system’s explanations achieved a clarity rating of 3.9, compared to 3.6 for human-made explanations, demonstrating its effectiveness in conveying information while incorporating humor;
- **Positive Response to Creative Content for Combating Misinformation:** 78.9% of participants responded positively to the use of creative and humorous content for addressing misinformation, highlighting the potential of this approach to engage and educate audiences.

All our work, including datasets and the fully implemented system, is available in our GitHub repository ¹.

1.4 Document Structure

The document is organized into seven chapters. Chapter 2 provides an overview of Deep Learning (DL) architectures and explores various aspects of humor. In chapter 3, we discuss systems capable of joke generation and fighting misinformation, along with moral and joke classifiers. Chapter 4 details our approach to the problem, outlining the goals of our Creative Explanation and the functionality of our system. In chapter 5, we cover the development process of the system, from the initial comparative study to select the model, to the implementation of all components that enabled the system to work. Chapter 6 presents the questionnaire used to evaluate the system, including an analysis of the results. Finally, chapter 7 summarizes the conclusions of our work and offers insights for future research.

¹ <https://github.com/hugo-ribeiro-36/Tese>

2

Background

Contents

2.1 Deep Learning Architectures	7
2.2 Computational Creativity	11
2.3 Humor	12

2.1 Deep Learning Architectures

In the realm of deep learning architectures, models emulate neural networks found in the human brain. They operate by processing information through interconnected layers of nodes. These architectures work by learning hierarchical representations, allowing them to perform diverse tasks.

2.1.1 Recurrent Neural Networks

Recurrent Neural Network (RNN) is a type of artificial neural network commonly used to model sequential data and capture patterns or dependencies within it. They have connections between nodes that form a directed graph along a temporal sequence, which allows them to maintain a hidden state or memory of past inputs, meaning they can capture information from previous inputs and take it into account to process new inputs [8]. This architecture makes them particularly suitable for tasks such as natural language processing, speech recognition, and time-series prediction.

However, traditional RNNs may face challenges in learning long-term dependencies due to issues like vanishing/exploding gradients. The vanishing gradient problem occurs when the gradients become extremely small as they are propagated backward through the layers of the network. To address these challenges, variations of RNNs have been developed, such as Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) Units.

2.1.1.A Long Short-Term Memory Networks

LSTM networks were designed to address the RNNs vanishing/exploding gradient problem by incorporating memory cells and gating mechanisms to control the flow of information [9].

LSTMs include several components:

- Cell state – These store information over time periods;
- Input gate – Determines which values are important and will be updated in the cell state;
- Forget gate – Determines what information from the cell state will be discarded;
- Output gate – Controls how much information of the cell passes to the output.

Their ability to mitigate the vanishing gradient problem makes LSTMs a powerful tool for learning from sequential data with long-range dependencies.

2.1.2 Sequence-to-Sequence Modeling

Sequence-to-Sequence (seq2seq) models are a type of neural network architecture used for tasks involving sequential data, especially in tasks where the input and output sequences are of different lengths.

The Seq2Seq model consists of two main components: an encoder and a decoder. The encoder takes in the input sequence and encodes it into a contextualized fixed-dimensional representation. The decoder receives the representation generated by the encoder in order to create the output sequence [10].

These models have been improved over time with variations like attention mechanisms, resulting in better performance in handling long sequences.

2.1.3 Attention Mechanism

Attention mechanisms assign a different importance to each part of the sequence when making a prediction. This way, the model can selectively attend to the most relevant information, improving its ability to generate accurate and contextually rich output sequences. Self-attention is a specific type of attention where each element within the sequence evaluates its relationship with all other elements in the same sequence. The introduction of self-attention is a breakthrough in handling sequential data, and it serves as the foundation for more advanced models like the Transformer.

2.1.4 Transformer

Transformers are the first sequence transduction models based entirely on attention in order to draw global dependencies between input and output [11].

First, since the Transformer does not inherently understand the order of the sequence, positional encodings are needed. They add to the input embeddings information about the order of input elements, which is essential in sequences but ignored by attention mechanisms. Following this, the input passes through the encoder, which consists of a stack of identical layers, where each one has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The Multi-Head Attention mechanism parallelizes the attention operations for different parts of the input and then the result is concatenated and passed to the position-wise fully connected feed-forward network followed by layer normalization. Then the sequence goes in the decoder, which is also composed by a stack of identical layers, but it divides each one into three sub-layers. The first two sub-layers are the same as in the encoder and the third one performs multi-head attention over the output of the encoder stack. It should be noted that the self-attention sub-layer in the decoder is modified in order to prevent positions from attending to subsequent positions, which prevents information leakage from future positions during training. The final decoder layer's output is transformed into probabilities using a softmax function, producing the model's prediction for each element in the output sequence.

One model built on the transformers architecture is the Generative Pre-Trained Transformer (GPT) [12]. GPT, that operates exclusively with the decoder component, is a generative model designed for text

generation, which is capable of generating coherent and contextually relevant text based on the input it receives. GPT goes through a pre-training phase where it learns from a diverse dataset. This allows the model to capture a wide range of language patterns and knowledge. As we will explore in the following sections, there are several strategies to further enhance the performance of models like this. As we will explore in the following section, there are several strategies to further enhance the performance of models like this.

2.1.4.A Prompt Engineering

A prompt, written in natural language, instructs the AI on the desired output, making clarity, meticulousness, and lack of ambiguity crucial. Imprecise language can lead to misunderstandings and inaccurate results, so prompts must include sufficient context, detailed instructions, and clear expectations. This can involve providing contextual details, explaining essential terminology, and stipulating the preferred style or demeanor of the final output. Incorporating examples of the expected behavior can further enhance clarity, especially in creative tasks, which is the case in this thesis.

Users interact with Language Model (LM) via prompts to describe tasks or provide inputs. Prompt engineering emerged as a technique to optimize these interactions, ensuring the model understands the task and expected behavior, thereby generating better outputs. Effective prompt engineering guides the model to produce useful outputs and iteratively refines the prompt based on the model's responses [13].

Prompts can be designed using various approaches:

- **Zero-shot:** No example is provided in the prompt and the model is presented only with the task description and the model must rely entirely on its pre-existing knowledge;
- **One-shot:** After the model is presented with the task description, one example is provided in the prompt to guide the model's behavior;
- **Few-shot:** Various examples are provided in the prompt after the model is presented with the task description. By providing multiple demonstrations, the model can better infer patterns in the desired behavior.

The iterative nature of prompt engineering means that prompts are continually refined to improve performance. Different formulations of a prompt, even if seemingly similar, can lead to different outputs, as models interpret variations based on their training data and context [13].

Overall, prompt engineering is a dynamic and iterative process that adapts prompts to align with both the model's training and the specific task requirements, enhancing the AI's ability to generate accurate and contextually appropriate responses.

2.1.4.B Fine-Tuning

The performance of a language model is highly dependent on the data it was trained on. Models trained with datasets tailored to the target task tend to show superior performance on that task. However, as models become more complex and their parameter counts increase, the cost and time required for training these models from scratch escalate significantly.

To mitigate these issues, the technique of fine-tuning has been introduced. Fine-tuning allows researchers to adapt a pre-trained language model for a specific task by further training it on a smaller, task-specific dataset. This method is beneficial as it avoids the need to start training from the ground up, leveraging the knowledge the model already possesses from its broad, initial training [14]. This approach streamlines the process, reducing both the time and computational resources required, while enhancing the model's ability to handle specific tasks.

When fine-tuning a model, the quality and preparation of the dataset are critical. Datasets should be carefully cleaned to remove unnecessary elements such as redundant spaces or repeated patterns, which can hinder the model's learning process. Additionally, ensuring that there are ample high-quality examples helps the model better capture the nuances required for the task. Poor dataset quality can lead to less reliable or overly deterministic outputs.

In summary, fine-tuning is a powerful tool for refining pre-trained models to excel in specialized areas, as long as the dataset is carefully curated and prepared.

2.1.4.C Hyperparameters

Hyperparameters are configuration settings used to control both the training and inference processes of a machine-learning model. Unlike model parameters, which are learned during training, hyperparameters are set before or after the training and play a crucial role in determining the model's performance and behavior [15].

There are three types of hyperparameters: those related to the network structure, the training algorithm, and the inference process.

A – Network Structure Hyperparameters:

- **Number of Hidden Layers:** These are the layers between the input and output layers, crucial for capturing complex patterns in the data;
- **Dropout Layer:** A regularization technique used to prevent overfitting by randomly dropping units during training;
- **Activation Function:** Introduces non-linearity into the model, enabling it to learn non-linear relationships.

B – Training Algorithm Hyperparameters:

- **Learning Rate:** Defines the speed at which the model learns by updating its parameters. A small learning rate ensures smooth convergence, while a high learning rate speeds up learning but risks not converging;
- **Number of Epochs:** The number of times the entire training dataset is passed through the model during training;
- **Batch Size:** The number of training samples used in one iteration to update the model parameters.

C – Inference Process Hyperparameters:

- **Temperature:** A value ranging from 0 to 1 that controls the randomness of the model's outputs. Higher values result in more random outputs, while lower values, such as 0.2, make the model's outputs more deterministic;
- **Top-p:** Also known as nucleus sampling, this hyperparameter considers the smallest set of top tokens whose cumulative probability exceeds a specified threshold p . For example, if $p = 0.1$, only the tokens whose combined probability mass exceeds 10% are considered. This allows for dynamic adjustment of the number of tokens based on their probabilities, ensuring that lower-probability tokens are included in sampling when the high-probability tokens are not sufficiently probable on their own.

2.2 Computational Creativity

The field of CC aims at building systems which display actions that unbiased observers would consider creative [16]. By modeling, reproducing, building and using computer programs, which show this kind of behavior. This field introduces concepts on how computers can independently generate works that have an emotional impact [7]. One possibility within this field is Computational Humor, which is the goal of this thesis, as we revolve mainly around humorous responses. Besides the creation of humor, there are a few relevant aspects related to the field of computational humor:

- **Humor Recognition** that deals with the task of deciding whether a given text is or is not humorous [17–19];
- **Humor Adaption Systems** are systems that adapt to a person's humor preference and act according to it [20, 21];

- **Computational Humor Evaluation** research focuses on how to assess the performance of a humor generating system [22, 23];
- **Computational Humor Data sets and Corpora** that deals with collected and published humor;
- **Humor Generation Systems** which are a particular type of text generation systems. There are two major approaches to build these type of systems: Template-based systems and Neural systems.

2.3 Humor

Humor is a multifaceted and deeply human characteristic that has captivated researchers across various disciplines. It is not only a source of entertainment but also a crucial component of social interaction and psychological well-being. The study of humor spans numerous theories and models, each aiming to unravel its complexities and understand its diverse manifestations in human behavior.

The upcoming sections will delve into specific theories and styles of humor and other aspects, highlighting key concepts such as incongruity, surprise, and ambiguity, and examining how these elements contribute to the humor experience. By exploring these theories, we gain deeper insights into the mechanisms behind humor and its various functions in our lives.

2.3.1 Humor theories

Humor is an intrinsically human trait, and it has long been a subject of investigation, leading to an abundance of different theories that explain verbally expressed humor [24]. While there is no consensus among all these theories, the most modern formal humor theories emphasize the importance of “incongruity” as an element of humor, where incongruity refers to the lack of coherence between the interpretation of two parts of a joke, the setup and the punchline [25–27]. This underscores humor as a form of cognitive activity, a problem solving task in which the listeners/readers engage with different sets of verbal representations [28].

In addition to incongruity, the Semantic Script Theory of Humor (SSTH), proposed by Victor Raskin [29], offers a comprehensive framework for understanding humor. SSTH posits that humor arises from the juxtaposition of two or more overlapping scripts—mental frameworks or scenarios that are typically incongruous or opposed in some way. This juxtaposition creates an unexpected effect, as the conflict between the scripts leads to a cognitive shift or surprise that triggers laughter. The theory highlights the importance of script opposition and resolution of incongruity in generating humor.

2.3.2 Humor styles

Beyond understanding what makes something humorous, researchers have delved into how humor functions in everyday life, leading to the development of various humor styles. These styles categorize the ways individuals use humor, which can significantly impact psychological and social well-being.

Martin et al. proposed four humor styles that serve as everyday functions of humor relevant for psychological and social well-being [30]. These styles are divided into two major groups corresponding to different functions of humor: adaptive, which includes Affiliative Humor, aimed at amusing others in a benign manner, and Self-enhancing Humor, characterized by a humorous attitude towards stressful events and adversities. The maladaptive group includes Aggressive Humor, which is used to downgrade other people through sarcasm and ridicule, and Self-defeating Humor, where individuals direct humor against themselves to amuse others.

Later building on Martin et al.'s four humor styles, Ruch, Heintz, et al. introduced eight comic styles which are fun, humor, nonsense, wit, irony, satire, sarcasm, and cynicism [31]. The aim was to complement the styles that already existed by investigating lower-level styles. They accomplished a narrower version than those introduced before, which allowed for a more specific distinction of humor-related styles.

Later, building on Martin et al.'s four humor styles, Ruch, Heintz, et al. introduced eight comic styles [31]. The aim was to complement the existing styles by investigating lower-level categories. They developed a more refined classification that allows for a specific distinction of humor-related styles. These styles include:

- **Fun:** Light-hearted amusement without deep intellectual engagement;
- **Humor:** Various forms of amusing expression, including jokes and anecdotes;
- **Nonsense:** Absurd or irrational humor that challenges conventional thinking;
- **Wit:** Clever wordplay or quick thinking in verbal expression;
- **Irony:** Conveying the opposite of what is expressed for humor or satire;
- **Satire:** Criticizing societal issues or human behavior through humor or exaggeration;
- **Sarcasm:** Mocking humor that uses irony to convey contempt or ridicule;
- **Cynicism:** Skeptical or pessimistic humor that questions sincerity or highlights flaws.

From these foundational studies, researchers made a conclusive effort to combine these two prominent theories, resulting in a taxonomy of nine different “styles” of humor. This not only advances our understanding of individual differences in humor but also provides a more refined way of exploring humor functions and expressions [32].

2.3.3 Relation between Humor and Language

As seen in the previous sections: 2.3.1 and 2.3.2, appreciating humor involves recognizing and resolving incongruity, which demands strong language skills to understand and interpret words within their sentence context. This raises an intriguing question: Are jokes funnier in one's native language?

This question was explored in a study conducted in 2017 [33]. In this study, which had a total of 160 participants, several questions were asked in order to examine the relationship between language comprehension and humor. The jokes presented were carefully chosen to exclude any cultural references and minimize language play. The study concluded:

- When jokes were translation equivalents and participants understood both, there was no significant difference in funniness between the native and foreign language versions;
- Ease of understanding was the primary factor influencing humor ratings in a foreign language, indicating that lower linguistic comprehension usually results in lower humor ratings;
- The authors propose a proficiency \times investment theory, suggesting factors that might make foreign language jokes funnier than those in one's native language. These factors include high proficiency levels and significant investment in the foreign language.

This theory has three main points:

- **Foreign language jokes will be experienced as less funny than native language jokes**
when: people have some proficiency deficiencies and lack investment in the foreign language;
- **Foreign language jokes will be experienced as equally funny to native language jokes**
when: people have no proficiency deficits or have a high level of investment in the foreign language (e.g., for a future job or career). This investment can enhance the pleasure of understanding the joke, making it funnier;
- **Foreign language jokes will be experienced as funnier than native language jokes**
when: people have high proficiency levels and a high level of investment in the foreign language.

It is noteworthy that the requirement for frame-shifting in humor underscores its complexity, revealing the formidable challenge for machines to acquire this skill, marking it as an AI-complete problem, presenting a challenging frontier for artificial intelligence research [24, 34].

3

Related Work

Contents

3.1 Joke Generation Systems	16
3.2 Misinformation Fighting Systems	26
3.3 Humor Classification Systems	28

This section reviews and critically evaluates the relevant literature that shapes the present research's context and trajectory. By delving into established theories and key findings, this review aims to establish a framework for comprehending the current state of the field.

3.1 Joke Generation Systems

In this section, our exploration begins with a comprehensive discussion on systems that generate humor. From template-based approaches to advanced neural systems, our investigation explores various technologies that play a role in the field of Computational Humor.

3.1.1 Template-based Systems

Template-based systems are a method of generating content using predefined templates, blocks of text with slots that can be filled with different variables [35]. It also requires the association of a schema - a framework that defines the relationships between the variables of a template [36]. As we are talking about systems for joke generation, these relationships are typically chosen to provoke incongruity, posing the primary challenge inherent in such systems: to fill in the slots of the template in order to make the whole phrase funny.

The first system generates jokes in the format: "I like my X like I like my Y, Z," where X and Y represent two nouns, and Z is an adjective or attribute [37]. For example, in the joke "I like my relationships like I like my source, open," X and Y are "relationships" and "source", respectively, while Z is the attribute "open". This type of joke relies on four key assumptions, which are illustrated in Figure 3.1.

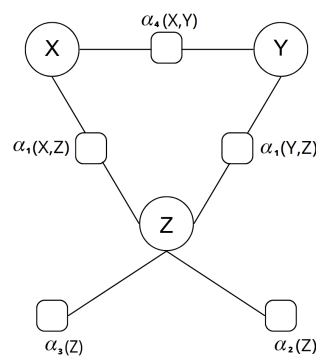


Figure 3.1: Template-base system proposed by Petrovic et al.

The first assumption comes from Formula α_1 (3.1), which is used to measure the similarity between nouns (X,Y) and attributes (Z). The lowercase variables within the formula represent random variables

assuming specific values. This calculation is grounded in the assumption that humor increases when attributes are frequently used to describe both nouns.

$$\alpha_1 = \frac{f(x, z)}{\sum_{x, z} f(x, z)} \quad \text{where} \quad f(x, z) = \text{Number of co-occurrences between } x \text{ and } z \text{ in a large corpus.} \quad (3.1)$$

Formula α_2 (3.2) leads us to the second assumption, that states that the less common the attribute is (divide by the number of occurrences of Z), the more likely it is to lead to surprise, thus making it funnier

$$\alpha_2 = \frac{1}{f(z)} \quad \text{where} \quad f(z) = \frac{1}{\text{Number of occurrences of } z}. \quad (3.2)$$

Expression α_3 (3.3) refers to the third assumption, which states that the more ambiguous the attribute is, the funnier the joke will be, so we want Z to have the lowest number of senses possible. This may seem counter-intuitive initially, but it occurs because a word with fewer meanings is more likely to be interpreted in multiple ways in different contexts. In contrast, a word with many meanings might be too versatile, reducing its comedic potential since it's already expected to be flexible.

$$\alpha_3 = \frac{1}{senses(z)} \quad \text{where} \quad senses(z) = \text{Number of different senses that attribute } z \text{ has.} \quad (3.3)$$

The last assumption claims that the more dissimilar the two nouns are, the funnier the generated joke will be and this is represented by formula α_4 (3.4)

$$\alpha_4 = \frac{1}{sim(x, y)} \quad \text{where} \quad sim(x, y) = \text{Similarity between } x \text{ and } y. \quad (3.4)$$

The function $sim(x, y)$ serves as a similarity metric, quantifying the similarity between the nouns x and y .

It should also be noticed that this system does not rely on labeled training data, and it instead relies on large quantities of unannotated data for these four formulas, using Google 2-grams [38], which analyzes pairs of consecutive words in a pre-existing dataset and Wordnet, a lexical database that groups words into sets and shows relationships between them [39]. To evaluate the system's performance, the jokes generated by it were compared against human-generated ones in a user study and were rated on a scale of 1-3, where 1 signified "funny", 2 denoted "somewhat funny", and 3 indicated "not funny". The study findings revealed a preference for human-generated jokes, with a higher percentage (33.1%) rated as "funny". However, the system demonstrated commendable results, achieving a mean rating of 2.5 compared to the human jokes mean of 2.1, and a notable 17% of the system's jokes were rated as "funny".

Yamane et al. [40] further developed the earlier system by incorporating new joke templates alongside the original ones, building upon the foundational work of Petrović and Matthews (2013) [37]. The overview of the enhanced system is depicted in Figure 3.2. Their approach introduces a dynamic expansion of third-person subjects, and incorporates fresh perspectives that encompass diverse relationships, ages, and interests. Examples such as “The baby likes his X like he likes his Y, Z” and “The middle-aged man likes his X like he likes his Y, Z” showcase these new templates, despite the author not providing examples for these templates, they consider they contribute to heightened humor variety. Additionally, Theory of Mind plays a crucial role in enhancing the depth and diversity of generated humor. Theory of Mind is the ability to understand what someone else might be thinking or feeling and to guess what actions they might take in a particular situation based on that understanding [41]. This cognitive element enhances the system’s capacity to craft jokes with a nuanced understanding of human behavior. Furthermore, Yamane et al. introduced a moral classifier within the system. This addition involves a selection mechanism that evaluates jokes based on moral judgments. This is achieved by training an LSTM on the Moral foundations twitter corpus [42]. The corpus is based on the moral foundations theory, which identifies five core moral factors or foundations found across cultures [43], and it has annotations for pairs of opposite moral foundations representing the interplay between the following pairs:

- **Care/Harm** – This foundation is related to the ability to feel empathy and compassion towards others, and to dislike the suffering of others. It underlies virtues such as kindness, gentleness, and nurturance. In terms of moral behavior, it promotes actions that protect others from harm and provide care and support;
- **Fairness/Cheating** – This foundation is rooted in the evolutionary process of reciprocal altruism, which promotes fair treatment, justice, and equity. It generates ideas of justice, rights, and autonomy. Virtues associated with this foundation include honesty, integrity, and trustworthiness. It discourages actions perceived as cheating or exploiting others;
- **Loyalty/Betrayal** – This foundation focuses on the importance of loyalty to one’s group, whether it be family, community, or nation. It underlies virtues such as patriotism, self-sacrifice, and solidarity. It emphasizes the value of group cohesion and the moral imperative to support and defend one’s group, while it condemns betrayal and disloyalty;
- **Authority/Subversion** – This foundation pertains to the respect for tradition, hierarchy, and legitimate authority. It underlies virtues associated with leadership, followership, and respect for societal structures. It emphasizes the importance of maintaining order and stability through deference to established authority and adherence to social norms;
- **Purity/Degradation** – This foundation is shaped by the psychology of disgust and the notion of contamination, both physical and moral. It underlies the idea that the body and soul can be dese-

crated by immoral activities and impurities. Virtues associated with this foundation include chastity, wholesomeness, and piety. It promotes behaviors that are seen as pure and condemns actions viewed as degrading or contaminating.

The objective is to select joke candidates based on diverse criteria, including the polarity of moral elements. The intensities of polarity and categories are assigned, and joke candidates are selected based on AND operations for each threshold. The thresholds ensure that the selected items meet specific sensitivity criteria for both polarity and categories.

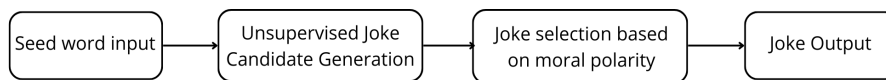


Figure 3.2: Template-based system proposed by Yamane et al.

The third system generates two types of jokes based on compound nouns in Wordnet [44]. This system introduces the concept of the “funniness” of a word, determined by its frequency in jokes compared to regular text. The underlying assumption is that words appearing more frequently in jokes are inherently funnier and lend themselves to the creation of humorous content.

To calculate the funniness of each word, the authors used two distinct corpora. The first contains 7,000 one-liner jokes, while the second consists of an equal number of randomly selected sentences from the British National Corpus [45]. The funniness metric is calculated using several criteria:

- If a word appears in both corpora, its funniness is determined by dividing its frequency in the joke corpus by its frequency in the non-joke corpus;
- If a word appears only in the joke corpus, its funniness is calculated as twice its frequency;
- If a word is exclusive to the non-joke corpus, its frequency is set to 0.1, and funniness is calculated as in the second case;
- Words absent from both corpora are assigned a funniness score of 0.

Additionally, there are also a few more specific criteria: a predefined list of stop words that have their funniness rated as 0, words that are frequent but lack humor (e.g., “do”), as well as inappropriate words are included in this list. Furthermore, capitalized words have their funniness divided by 5 to avoid jokes targeting specific people or locations. Conversely, dirty words, often linked to adult humor, have their funniness multiplied by 10.

The system has the capability to produce two distinct types of jokes as detailed in Figure 3.3, and it generates numerous possible jokes based on the information Wordnet provides. Leveraging the concept of the funniness of a word, the system calculates the overall funniness of a sentence from all potential joke candidates. The final output is determined by selecting the joke with the highest funniness score.

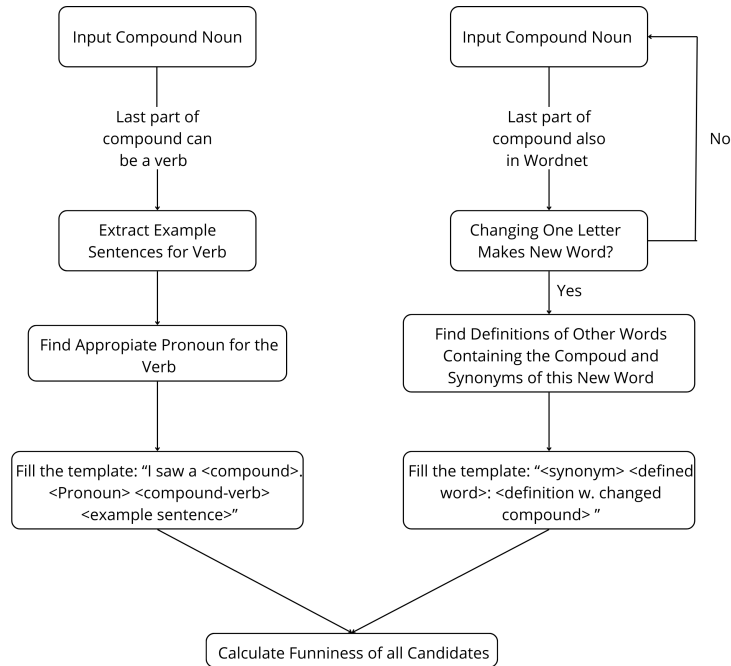


Figure 3.3: Types of joke production of the template-based system proposed by Sjobergh et al.

The final template-based system, introduced by Labutov and Lipson, distinguished itself as the pioneer in explicitly incorporating a linguistic humor theory [46]. In this approach, the system extracts simple humorous scripts from a semantic network, specifically ConceptNET, a vast network of interconnected concepts and the relationships between them. This method involves a targeted search for dual scripts that maximize both overlap and incongruity aligning with the principles of the SSTH [47] - the humor theory it adheres to. The proposed approach involves three stages:

1. Building a Script Model (SM), that captures likely transitions between concepts in a sequence;
2. Generating Feasible Circuits, by using the script model to find paths from a starting point, producing a set of ranked scripts;
3. Converting to Surface Form, the ranked scripts are transformed into natural language templates (Question/Answer format) by aligning concepts. The alignment prioritizes incongruity for humor.

A script is represented as a series of connections between concepts, forming a first-order Markov chain. Starting from a seed concept, a depth-first search is performed, exploring all paths that end at the same node. These paths are potential script pairs, but not all of them result in meaningful content, so they need to be filtered. To do this, they assign binary labels (feasible or infeasible) to each potential script. The system is then trained and the learned transition probabilities help capture basic stories with events, consequences, and qualifiers like certainty, time, size, and location. In essence, the script

model learns from examples to create meaningful and coherent scripts by considering the relationships between concepts in a chain.

The script model captures meaningful transitions within a single script but does not consider overlap and incongruity between different scripts. So a modified spreading activation method, technique used to simulate the flow of information or influence through a network, is used to find semantic circuits while maximizing their semantic overlap. Starting from a user-specified root concept, activation radiates along outgoing edges, considering transition probabilities and penalizing for long scripts. Activation continues until weights, that depend on the transition probabilities involving a node, fall below a threshold. Nodes are ranked by activation weight. For each highly ranked node, a set of paths (scripts) is scored based on their length and likelihood, favoring those with semantic overlap. The highest-scoring scripts represent feasible sequences with substantial semantic overlap.

The task involves creating script pairs and concepts that follow a template while maximizing inter-script incongruity. This incongruity is measured using hierarchical clustering on ConceptNet. This clustering method organizes related concepts into groups. The template has a structured question-answer format with carefully selected concepts as seen in Figure 3.4. The method strategically selects two concepts ($c1$ and $c2$) from different scripts ($S1$ or $S2$) but the same cluster for the question, while the answer concept ($c3$) is chosen from one script and a different cluster. This setup and resolution approach create a punchline effect. In this context, the question deliberately establishes two parallel and compatible scripts, while the answer serves as a catalyst for a change in the underlying script. Let's break down an example to further explain this concept, using the joke: "Why does the [priest] kneel in [church]? Because the [priest] wants to [propose woman]." In this example, the system effectively utilizes two scripts that share the common concept of a "priest", aligning it with a religious setting, represented by "church". The question sets up an expectation of a religious or spiritual action, as kneeling in church is commonly associated with prayer or worship. However, the punchline introduces an unexpected twist by suggesting that the priest has romantic intentions, indicated by "propose woman". This shift from a religious context to a romantic one creates incongruity, eliciting humor by surprising the audience and subverting their expectations. The alignment of concepts from different clusters—religion and romance—highlights how the method maximizes inter-script incongruity while maintaining coherence in the setup.

3.1.2 Neural Systems

The latest trend in computational joke generation involves training neural networks to generate humorous texts as output [35]. Yu et al. [48] proposed a system that uses a seq2seq architecture to generate puns, without relying on a pun corpus for training. Instead, the system is trained using a conventional text corpus. It then takes two senses of a target word as input and produces a pun, where both senses of the word are contextually appropriate. For instance, consider the word "problem", which can refer either

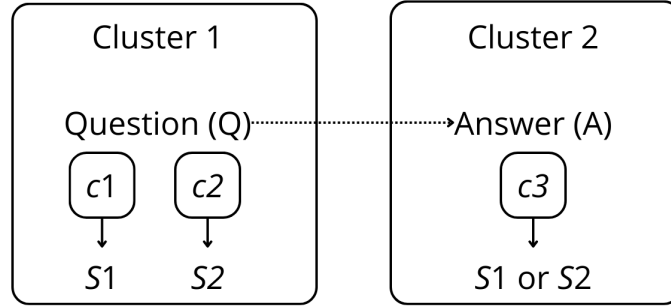


Figure 3.4: Template for the jokes of the system proposed by Labutov et al.

to a source of difficulty or a question posed for consideration. In the joke “Math teachers have lots of problems”, both senses of the word are meaningfully integrated into the punchline.

Initially, an encoder-decoder framework is employed to train a conditional language model capable of generating sentences that incorporate one of the provided senses of the target word. An LSTM serves as the encoder, to map the input sequence into a fixed-dimension vector. Given that conventional seq2seq models cannot ensure the appearance of the target word in the output, and considering its high importance in the creation of the pun, the authors use an asynchronous forward/backward generation model [49], that guarantees that the word appears in the output of seq2seq models. Since the idea is to generate a sentence containing one of the word senses, the approach involves treating different senses of the same word as independent new pseudo-words. The model is subsequently trained using a corpus with labeled senses, enabling it to generate sentences tailored to each distinct word sense.

In the decoding stage a joint beam search is used, which is an extended version of the original beam search. Unlike the traditional version, which selects the most promising nodes for a single sequence, the joint beam search extends its scope to consider the most promising nodes for multiple sequences. This means that during the decoding stage, the algorithm explores potential sequences for distinct word senses in parallel. The selection of promising nodes at each level of the search graph is not only based on individual senses, but also takes into account a joint score distribution, allowing for the generation of sentences suitable for both senses.

To improve the generated sentences, a key enhancement involves incorporating relevant keywords that evoke specific senses of the target word. To identify these keywords, word association norms are established in the corpus using pointwise mutual information (PMI). This metric compares the probability of co-occurrence between word pairs with the probability of observing each one independently. Positive PMI scores indicate a higher likelihood of words occurring together [50]. The selection of keywords involved extracting the top k associative words with positive PMI scores, thereby ensuring that the chosen keywords are most indicative of the semantic nuances.

The next system was proposed by Ren et al. [51] and it uses an encoder-decoder architecture to generate a short joke related to the topic that the user provides. To start with, the encoder is trained with data from two different sources. The first one contains short jokes written by Conan O'Brien, while the second one consists of online news data from the Mashable website. Then a Part-of-Speech Tagger (POS Tagger), which is a natural language processing tool that assigns tags (such as noun, verb, adjective, etc.) to individual words within a text, is used to extract the nouns from each training samples. This is done under the assumption that the nouns are like keywords that can represent the topic of the joke, given that the goal is to generate jokes based on topic words given by the user. Then, GloVe embeddings, a type of word embedding technique that captures semantic and syntactic features, are used to encode the proper nouns. With this process, the encoder learns to convert input sequences into fixed-dimensional vectors, that are then passed as the initial hidden state of the decoder.

In the decoder, an LSTM network with an attention mechanism is used to generate the joke containing the topic words. Moreover, the system accommodates user preferences by allowing them to provide multiple options, such as: specify the topic of the joke, what words the joke will start with and the exact number of words that the system will output.

While many joke production systems suffer from limitations such as the absence of explicit humor algorithms and reliance on a singular joke production mechanism, Joe Toplyn addressed these shortcomings with the Witscript3 [52]. This system introduces the Basic Joke-Writing Algorithm, a comprehensive five-step approach for crafting a three-part joke, aimed at producing a more consistent and diverse range of humor. The five steps are:

1. **Choose a topic:** a good joke topic is a concise sentence that is likely to grab the audience's attention;
2. **Identify two topic handles:** select the two most attention-grabbing words or phrases within the chosen topic;
3. **Create associations:** generate associations, what the audience may think of when considering the subject, related to the two topic handles;
4. **Develop the punch line:** devise a word or phrase that induces laughter by cleverly connecting associations of the two topic handles in an unexpected manner;
5. **Establish an angle:** formulate a natural-sounding word sequence, the angle, that seamlessly links the topic to the punchline, enhancing the overall comedic impact.

The system initiates the execution of the algorithm by invoking OpenAI's GPT-3 [53]. It is important to highlight that the model is used in its pre-trained state, without undergoing any fine-tuning. The algorithm consists of five distinct steps carried out independently. This implies that the system performs individual

calls to the model, using the output of each step as the input for the subsequent one. By breaking down the process into steps, it establishes a transparent workflow, where a step in the algorithm corresponds to a specific stage of the task, allowing for a clear understanding of how the model processes information sequentially, and enabling the system to enhance its controllability and explainability. This procedural approach is known as prompt chaining [54].

To assess the effectiveness of the system in generating jokes, a user study was conducted. Participants were tasked with rating jokes from three distinct sources: human-made sentences, jokes generated by a simple joke generator serving as baseline and jokes crafted by Witscript3. The study revealed promising results, with Witscript3 achieving a mean rating of 2.36 on a scale ranging from 1 to 4. This surpassed the score of human-made sentences, which stood at 1.86. It is important to note that the human sentences used in the study were not crafted with the aim to be humorous. Consequently, this aspect warrants caution when interpreting the comparison, as the absence of humor-oriented intent in the human-generated sentences may introduce a limitation in assessing the model's performance in generating humorous content.

The final neural system, created by Oskar Niemenoja [55], produces homographic puns. To achieve this, the author firstly looks for words that follow four principles:

- The words used are standard and easily understood in English.
- The words have a shared element, referred to as the “connector”, present at the end of the first word and the beginning of the second;
- The words make sense even if we remove this common part (connector);
- The connector is specifically a non-English word, and it is required to be of a certain length, ensuring its distinctiveness.

An example of a pair of words that follows these principles is “maple” and “lessons”, where the connector is “/es”. Even upon removing it, the resulting words, “map” and “sons”, remain linguistically sensible.

To look for them, one option would be to go through all of the words in the English language and check if any of those have overlapping substrings with any of the other words. The used approach implements Tries, a type of search tree where the search strings are sequences of characters. Each character in a word leads to the creation of subtrees, and the final character of a word lands on a leaf that stores that value. So, if a node is a word in itself and contains children, the node is stored in a list. With this process, the list of homographic words to use is ready, resulting in around 10000 wordplays. Then, a pre-trained transformer model, OpenAI's GPT-3 [53], is loaded, receiving a prompt that contains one of the randomly chosen wordplays.

3.1.3 Hybrid Systems

This type of systems shares properties of both aforementioned sections. Valitutti et al. [56] created a system that takes as input a short English text message and performs a single word substitution with the objective of producing a funny response. This system is considered a hybrid because it combines neural methods with predefined constraints or templates to generate humor. While it leverages neural networks to model the language and select substitute words, it simultaneously adheres to specific rules and restrictions.

The substitution is based on three main constraints, where each one carries a distinct weight in shaping the overall funniness of the output, and together, they mutually support each other:

- The substitute word has to be similar to the original word, phonetically or orthographically, thus turning the original text into a pun. Two words are deemed orthographically similar if one can be transformed into the other one with a single operation such as deletion, addition or replacement of a character. On the other hand, words are regarded phonetically similar if their phonetic transcriptions exhibit orthographic similarity;
- The substitute word needs to be a taboo word or used in a taboo expression. According to Jay et al.(2008), “Taboo words have uniquely strong connotative meanings” [57]. There are two distinct categories of taboo words: Connotational taboo words, where the taboo is inherent in the utterance itself, and Taboo-inducing words, which have the potential to elicit taboo meanings when used as substitutes in a short text;
- The substitute word should replace the original one at the end of the text or in a place that is statistically consistent with the neighboring words. This approach increases local coherence, due to the fact that if the substitute word fails to form a coherent or plausible compound with its immediate predecessor or successor, then the resulting text is likely to lack clarity or coherence;

A user study was conducted to assess the system’s performance, with funniness being the primary criterion for evaluation. The obtained results were not good, yielding a mean funniness rating of 0.55 on a scale of 0 to 4. The responses, therefore, did not meet the desired humorous standard and could not be characterized as genuinely funny.

Table 3.1 provides a concise overview, summarizing the in-depth analyses presented earlier. It encapsulates information on the training datasets used by the various joke generation systems, accompanied by an example joke produced by each system.

3.2 Misinformation Fighting Systems

In this section, we explore the key components and challenges within the realm of misinformation fighting systems, shedding light on their essential role in maintaining the reliability and credibility of information in the digital landscape.

He et al. [58] introduced a reinforcement learning-based framework known as MisinfoCorrect. This system is designed to generate counter-misinformation responses. Operating on a foundation of reinforcement learning, it employs a policy network in conjunction with a transformer-based model adapted from GPT-2 [59]. When confronted with a misinformation post, the system generates a response that exhibits several desirable properties. These properties include:

- **Refuting** – the response explicitly disputes the misinformation to correct the spreader and ease the impact of misinformation;
- **Evidence** – the response incorporates supporting evidence, increasing its effectiveness in refuting misleading claims and potentially reducing the belief in misinformation;
- **Politeness** – the response maintains a polite tone to prevent potential backlash. Politeness increases the likelihood that the spreader will consider the accurate information more agreeably;
- **Fluency** – the generated text is fluid and expressive, ensuring a natural comprehension and engagement for readers;
- **Relevance** – the response is closely tied to the misinformation post.

The authors confronted a significant challenge, as there were no pre-existing datasets containing pairs of annotated misinformation posts and corresponding counter responses. To tackle this issue, they developed two novel datasets. The first encompasses authentic social media counter-responses, capturing real-world reactions to misinformation. The second dataset involves crowdsourced counter-responses, providing a controlled environment for gathering additional annotated data.

Using these datasets as a foundation, the authors proceeded to establish a Reinforcement Learning (RL) agent, characterized by four key components:

- **State** – In this reinforcement learning framework, the state (s) encapsulates the essence of the misinformation post (m). The objective is to use the input m to elevate the quality of the counter-response text, ensuring its relevance to the claims made in the misinformation;
- **Action** – Given the state s , the agent initiates the generation of a candidate counter-response \hat{c} . The text generator is represented as g , and the action $a = g(s)$;

- **Policy** – The policy is based on the transformer language model, specifically employing masked multi-head self-attention layers on GPT-2. In this framework, the input consists of an encoded representation of the state s , and output corresponds to the action a . The generation task is formulated as a language modeling problem, generating \hat{c} that maximizes the conditional probability $p(\hat{c}|m)$;
- **Reward** – The reward system incorporates novel functions designed to encourage the generated response to exhibit the above mentioned properties. All rewards are quantified with the use of a classifier that measures each characteristic.

Then, in the training phase the pre-trained weights of DialoGPT [60] are used to initialize the weights in the transformer-based GPT-2 [59]. The training approach for the Reinforcement Learning (RL) agent involves incrementing rewards gradually. This approach follows standard reinforcement learning principles, guiding the RL agent through a trial-and-error learning process based on positive feedback. The system undergoes training using both datasets, and upon analysing the results, the authors reach the conclusion that responses generated exhibit significantly higher quality when utilizing the crowd-sourced dataset. This underscores the substantial impact of dataset quality on the model's ability to yield superior results.

André Santos put forward a system that generates creative explanations for fake news through the use of poetry [61]. The system takes a condensed fake news explanation and transforms it to create poems, drawing inspiration from the content encapsulated in its explanation. The creative explanations are expected to meet specific criteria, ensuring they are informative, creative, accurate, appealing, and balanced.

In the quest to select the most suitable model for the task at hand, a comparison study was conducted, considering various models with different sizes and parameters. The models were prompted to generate poems, using a text sequence as input. The evaluation encompassed both quantitative and qualitative aspects. Quantitatively, the meter, feet, and rhyme scheme were analyzed using Poesy [62], a tool that annotates and generates a tabular summary of the poem. Meanwhile, qualitative evaluation relied on human assessment to verify whether the poems effectively conveyed the explanation and were easily readable. After careful examination of the results from both evaluations, GPT-3 DaVinci was the chosen model for training.

Following the model selection, a fine-tuning process began, using three distinct datasets. Consequently, three separate fine-tuned models were created. Among the datasets used, one consisted of annotated pairs featuring summaries of Fake News explanations and their corresponding poems. Another dataset presented pairs of claims matched with humorous poems, while the last dataset was a subset of the latter, specifically concentrating on the selection of quatrains.

To assess the performance of each model, a testing dataset comprising of twenty randomly selected

entries from the first dataset was used. Through a combination of quantitative and qualitative evaluations, it was determined that the poems generated by the model fine-tuned with the first dataset exhibited the higher resemblance to those crafted by humans, making it the one to choose. Subsequently, a questionnaire was designed in order to perform a comparison between the poems generated by the model and those composed by humans. The survey revealed that although the human made poems were preferred in terms of liking, humor and creativity, the model generated poems that were preferred in terms of explainability.

3.3 Humor Classification Systems

In this section, we delve into systems capable of humor classification, meaning that given a sentence, the system can determine whether it is a joke or not and, more importantly, attribute a score to it.

The first system identifies humor in texts using a widely recognized linguistic theory known as the incongruity theory of humor [63]. The architecture of the system, illustrated in Image 3.5, outlines the steps that follow the initial pre-processing of the input data, which ensures it is clean and well-structured. The data cleaning process involves the following steps:

- **Expanding contractions:** All contractions are replaced with their extended version. For example, “can’t” is turned into “cannot”;
- **Cleaning punctuation marks:** Punctuation marks were separated from words to achieve cleaner sentences. So if text has “Hello (World)!” , the result would be “Hello (World) !”;
- **Cleaning special characters:** Special characters are replaced by their alias. “Alpha” instead of “ α ”, for example.

Next, the system processes each sentence individually, tokenizing it using BERT sentence embeddings. These embeddings are then fed into parallel layers in a neural network to extract mid-level features (e.g., context, type of sentence) for each sentence. Simultaneously, the entire text is processed in another layer to extract features of the whole text. This is important because, in humorous texts, individual sentences might appear normal or non-humorous when isolated; thus, analyzing the text as a whole is crucial.

Finally, three sequential layers combine the outputs of the previous parallel layers to produce the final output. These final layers determine the congruity of the sentences and classify the text as humorous or not. The system was trained on 200,000 labeled short texts, with an equal number of humorous and non-humorous texts. It achieved high results, managing to attain an F1-score of 98.2%.

The second system uses GPT-4 to assess the funniness of a given joke [64]. GPT-4 has the capability of performing a comparative evaluation of textual outputs [65], enabling it to automatically evaluate

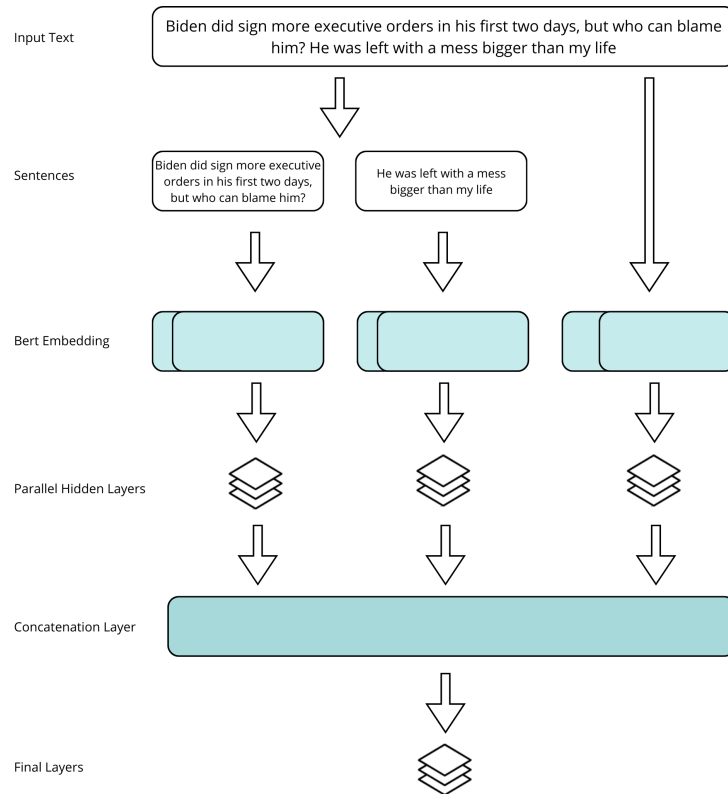


Figure 3.5: Architecture of a humor classifier based on a popular linguistic theory of humor.

creative tasks like jokes. This approach leverages the “System description” feature, where GPT-4 assumes different roles or personas based on provided prompts. This feature is used to describe various humor types, aiming to imitate human evaluators. Seven different system descriptions were used to simulate human responses. Two of them served as baselines: one with no system description and one with a naive description (e.g., “You are an humor expert”). Four were based on the humor styles proposed by Martin et al. [30] (affiliative, self-enhancing, aggressive and self-defeating) to encompass all humor styles. The final system description was generated using a few-shot prompt with 200 jokes and their ratings. This unique system description can potentially achieve the same accuracy and performance as a few-shot approach but at a lower cost, since there is no need to include a many examples in each inference.

In addition to system evaluation, five different prompts were tested:

- A few-shot prompt with 3 examples about how to score jokes from a scale of 1-5;
- A prompt that provides an explanation of how to score the jokes;
- A prompt that explains how to score jokes using examples and explanations, in other words a combination of the first two;

- A prompt that uses examples and explanation with a few additional examples from Additional Calibrating Examples;
- A baseline prompt with no examples or explanations.

To validate the results, a dataset of joke pairs with human ratings was used. Instead of comparing the system's ratings to human ratings directly, the ranking of jokes was compared to avoid scaling issues and focus on behavioral alignment. The Spearman correlation coefficient was used to measure the correlation, ranging from -1 (negative relationship) to 1 (positive relationship). With the generated system description and the fourth prompt, the system achieved a weak but positive correlation with human judges.

Type of joke/System name	Authors	Class	Dataset/Corpora/ Knowledge Graph	Input	Example
Unsupervised joke generation from big data	Petrović and Matthews	Template-based	Google n-gram data and WordNet	Not Applicable	I like my boys like I like my sectors, bad
Ambiguous Compound Joke Generator	Sjöbergh and Araki	Template-based	British National Corpus and WordNet	Not Applicable	I saw a Markov process. It (the Markov) processed cheese.
Humor as Circuits in in Semantic Networks	Labutov and Lipson	Template-based	ConceptNet	Not Applicable	Why is the computer in hospital? Because the computer has virus.
Neural Approach to Pun Generation	Yu, et al.	Neural	English Wikipedia corpus	problem: 1) a source of difficulty; 2) a question raised for consideration or solution.	Math teachers have lots of problems.
Neural Joke Generation	Ren, et al.	Neural	Short Jokes by Conan O'Brien and news from Mashable website	Los Angeles; Trump	According to a new study , the governor of film welcome the leading actor of Los Angeles area , Donald Trump.
Witscript3	Joe Toplyn	Neural	Pre-trained LLM	Germany has given animals legal rights in their constitution.	If animals have legal rights, does that mean I can sue my neighbor's dog for barking?
Procedural pun generation	Niemenoja	Neural	Pre-trained LLM	Write a joke containing the words "potent angle" and "pot entangle"	What do you call a potent angle that's also good at entangling? A pot-entangle!
Humorous SMS generation	Valitutti, et al.	Hybrid	NUS SMS Corpus	Remember to get the phone book from that person.	Remember to get the phone book from fat person.

Table 3.1: Joke generation systems reviewed.

4

Problem Approach

Contents

4.1 Creative Explanations	33
4.2 Approach	33

This chapter explores creative explanations for fake news, detailing the necessary requirements. We then present our methodology for developing our system.

4.1 Creative Explanations

Creative Explanations for Fake News aim to creatively clarify why a piece of information was assigned a particular label. These explanations are challenging to craft, especially given that most efforts to counter misinformation rely on textual justifications. These creative explanations need to be engaging and depart from the conventional approach of lengthy, jargon-filled textual explanations that are often disengaging and difficult to understand. However, they must still convey precise and accurate information without manipulating it, as misinformation does. As mentioned in Section 1.2, these explanations need to be: Creative, Enlightening, Precise and Appealing. With these characteristics, we aim to fight the rapid spread of misinformation effectively.

To develop these innovative explanations, we use online information and news articles from PolitiFact¹, a fact-checking website launched in 2007. For each news article they fact-check they provide:

- The claim to be debunked;
- A label classifying the claim, which can be one of six categories:
 - **True** – The claim is accurate and complete;
 - **Mostly True** – The claim is accurate but needs additional elucidation or information;
 - **Half True** – The claim is partially accurate but omits details or takes things out of context;
 - **Mostly False** – The claim has some truth but hides critical information;
 - **False** – The claim is not accurate;
 - **Pants on Fire** – The claim is not accurate and is ridiculously false;
- A detailed textual explanation of why the claim received that label;
- A section called “If Your Time is Short”, providing a brief explanation of why the claim has that label.

By leveraging this structured information, which we feed as input to our system, we hope to create more engaging and effective explanations to counter misinformation.

4.2 Approach

In this section, we present the comprehensive design of our system, present in Image 4.1, detailing the rationale behind each decision. Our goal was to create a system that effectively generates jokes with the requirements mentioned before. This involves several critical steps, from model selection to fine-tuning and evaluation.

¹ <https://www.politifact.com/>

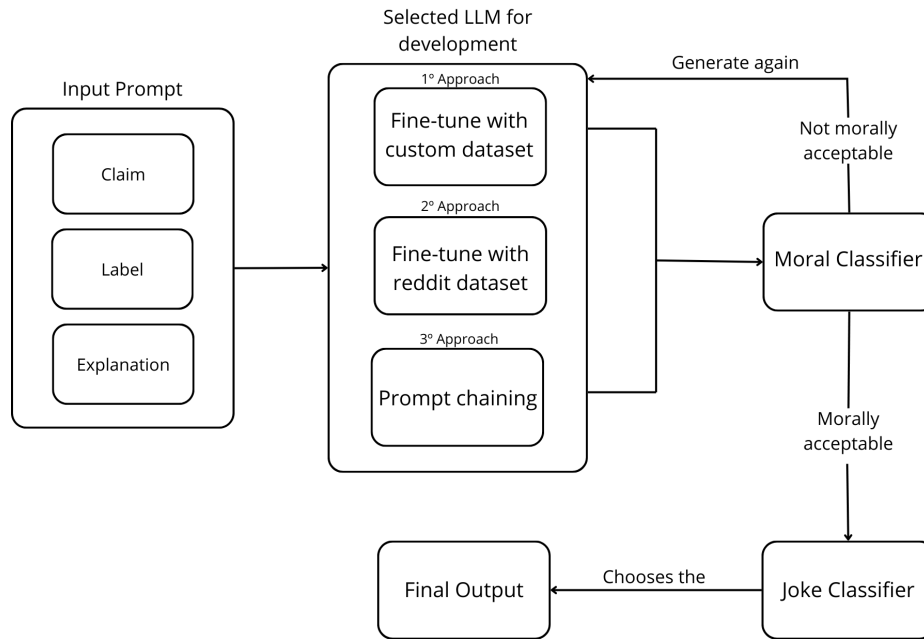


Figure 4.1: Architecture of our joke generation system.

4.2.1 Humor Generation Approaches Selection

In section 3.1, we reviewed several humor generation systems (Table 3.1), including template-based, neural, and hybrid approaches. Each system offers unique advantages and limitations:

- **Template-based systems**, such as those proposed by Petrovic and Matthews or Sjöberg and Araki, are highly structured and efficient for certain types of jokes. However, their fixed nature makes them unsuitable for handling dynamic and unpredictable content, such as fake news, where adaptability is crucial;
- **Hybrid systems**, which combine templates with neural methods, offer more flexibility compared to purely template-based systems but still depend on static patterns. While they are effective in certain cases, their potential to generate varied and context-specific humor remains limited when dealing with diverse topics like fake news;
- **Neural systems**, including those by Ren et al. and Yu et al., provide a more flexible solution. These models, trained on large datasets, are capable of generating contextually appropriate and diverse content, allowing for greater adaptability across multiple topics without being constrained by rigid templates.

Given these observations, we concluded that neural systems offer the greatest potential for generating the kind of humorous yet explainable content required by our system. While neural models may

require large datasets and careful fine-tuning to produce high-quality jokes, their adaptability makes them the most suitable choice for our goals.

Since no language models have been specifically designed for humour generation, limited information was available regarding their performance in joke generation. To address this gap, we conducted a comparative study of various models, described in detail in Section 5.1.1. The goal was to identify the most suitable model to be used in our system. Our evaluation focused on several key criteria, including the models' ability to understand context, generate coherent and relevant content, and their potential to produce humorous output.

4.2.2 System Components Overview

After selecting the best Large Language Model (LLM), we implemented three different approaches using it:

- Applying an Humour Algorithm with prompt chaining based on one of the systems we analyzed;
- Fine-tuning the selected model with a dataset containing jokes;
- Fine-tuning the selected model with a custom dataset created specifically for this task.

Each of these approaches is detailed in Section 5.2 and beyond. Furthermore, given that the goal of this system is to craft jokes to combat misinformation without insulting or degrading anyone, we implemented a moral classifier to filter the outputs from each approach. This classifier ensures that none of the generated jokes are offensive, which is a serious concern when aiming to fight misinformation.

Initially, our plan was to perform a comparison study between the three approaches and use the best one for the final system. However, we later decided to use a joke classifier to select the best joke from the outputs of all three approaches. This new strategy eliminated the need for a comparison study, allowing us to use all three approaches simultaneously in the final system.

4.2.3 Result Analysis

To evaluate the final system, we conducted a user study designed to assess its performance. Participants were asked a series of questions to provide unbiased feedback on various aspects of the generated jokes. The questionnaire aimed to determine whether the jokes could pass as human-made, whether they were funny, whether they effectively explained the news articles, amongst other important topics covered in the survey.

While designing the study, we considered several concerns to ensure the validity of the results. One major concern was avoiding participant bias; to address this, we ensured that all jokes were presented anonymously, with no indication of whether they were generated by our system or written by humans.

Another concern was the clarity of the questions—ambiguity could lead to misinterpretation, so we carefully worded the survey questions to be as straightforward and specific as possible. Additionally, we were mindful of the diverse backgrounds and humor preferences of participants, and we aimed to gather a broad range of opinions.

5

Development

Contents

5.1	Comparison Study	38
5.2	Joke generation Approaches	44
5.3	Moral Classifier	53
5.4	Joke Classifier	58
5.5	System Integration and Assembly	62

In this chapter, we detail the development process of our system, focusing on the various components that contribute to its functionality. This includes a comprehensive comparison study, an exploration of our joke generation approaches, and the design and implementation of the moral and joke classifiers. Each section provides an in-depth look at the methodologies, models, and techniques used, along with the evaluation processes that guided our choices.

5.1 Comparison Study

As mentioned before, there are several language models available, needing a comparison study to determine the most suitable one for further development. For the study, each model was tasked with crafting a joke given a claim, a label and an explanation. All models used were in their base state without any type of fine-tuning, and when possible, their hyperparameters would also be the same. This approach ensured that we tested all models in their original form.

Using this setup, we generated jokes for five different news articles, which were then used for our survey. To evaluate the results, we used a quantitative approach, rating each joke on three criteria: whether it qualified as a joke, its humor value (rated on a scale of 1 to 5), and its effectiveness in explaining the article's assigned label (also rated on a scale of 1 to 5).

5.1.1 Model List

Each model tested needed to meet two main criteria: it had to accept a text input and generate a joke, and it had to generate content in English. Since there is no specific model designed for joke generation, we focused on transformer-based models. The following models were selected for evaluation:

- GPT-2 – Developed by OpenAI, GPT-2 is a large-scale transformer model with up to 1.5 billion parameters. It marked a significant step in natural language processing (NLP) by generating coherent and contextually relevant text from prompts;
- GPT-3 – An evolution of GPT-2, GPT-3 boasts an impressive 175 billion parameters. GPT-3 includes multiple models, with notable ones such as “da-vinci-002” and “gpt-3.5-turbo-0125”, the latter being one of the best-performing models of its type;
- GPT-Neo – GPT-Neo is an open-source alternative to OpenAI's GPT-3, developed by EleutherAI. It aims to replicate the performance of GPT-3 with freely available models, making advanced text generation more accessible;
- GPT-4 – The latest in the GPT series, GPT-4 further improves upon its predecessors with enhanced capabilities and a deeper understanding of language nuances;

- T5 – The Text-to-Text Transfer Transformer (T5) by Google frames all NLP tasks as a text-to-text problem. The T5 model comes in various sizes, with the largest, T5-11B, containing 11 billion parameters. It is known for its flexibility and effectiveness across different NLP applications;
- Gemini – Gemini is a versatile transformer model designed for a broad range of text generation tasks, developed by Google DeepMind.

5.1.2 Development and Results

Once the conditions were established, we proceeded to generate jokes using each of the selected models and began developing our survey. During the testing phase, however, we encountered an issue with the Gemini model: its API was inaccessible in our country. This limitation forced us to exclude Gemini from our study and instead concentrate on evaluating the other five models. For each of these models, we tested different available sizes, ensuring a thorough comparison across different configurations.

The survey included eight participants—seven male and one female—with education levels ranging from Bachelor's to Master's degrees. For each generated joke, participants were asked three questions:

- Do you consider this to be a joke?;
- Indicate how funny the response is on a scale from 'Boring'(1) to 'Very Funny'(5);
- Assess the explainability of each response on a scale from 'Not explainable'(1) to 'Highly explainable'(5).

With this in mind, we now present more specific information about the testing and the results for each model.

5.1.2.A GPT-2

We first examine the results obtained by GPT-2. We used two different GPT-2 architectures: the base GPT-2 model with 124M parameters and GPT-2 Large with 774M parameters. These models were accessed via the HuggingFace library, and the prompt was adapted to generate jokes for our study.

The survey results for these models are summarized in Table 5.1.

Table 5.1: Survey Results for GPT-2 and GPT-2 Large

Model	Is it a joke? (Yes %)	Humor Rating (1-5)	Explanation Quality (1-5)
GPT-2	12.5%	1.60	1.63
GPT-2 Large	15.0%	1.68	1.7

As the table indicates, the performance of GPT-2 in generating humor was notably poor. Only 12.5% of the participants considered the outputs from the base GPT-2 model to be jokes. When we scaled up to GPT-2 Large, the percentage increased slightly to 15.0%, suggesting a marginal improvement with the larger model. However, this difference is minimal and does not signify a substantial enhancement in the model's joke recognition capability.

The humor rating, which was assessed on a scale from 1 to 5, also reflects this trend. The base GPT-2 model received an average humor rating of 1.60, indicating that the jokes were largely perceived as unfunny. The GPT-2 Large model fared only slightly better with a humor rating of 1.68. This suggests that increasing the model size had a negligible effect on the perceived funniness of the generated jokes.

Similarly, the explainability of the jokes, rated on a scale from 1 (Not explainable) to 5 (Highly explainable), showed poor results. The base model scored an average of 1.63, while the larger model scored 1.7. These low scores imply that participants found the jokes not only unfunny but also not capable of explaining why an article has a particular label, which is crucial for our study.

Overall, the results demonstrate that neither the base GPT-2 nor the GPT-2 Large models were effective in generating humorous content. The slight improvements observed with the larger model indicate that merely increasing the number of parameters is insufficient to significantly enhance the humor generation capability.

5.1.2.B GPT-Neo

The next model tested was GPT-Neo, for which we evaluated two of its architectures: GPT-Neo 1.7B and GPT-Neo 2.7B. Similar to the process with GPT-2, these models were assessed through their HuggingFace libraries, where we adjusted the prompts to generate jokes for our study.

Table 5.2: Survey Results for GPT-Neo 1.7B and GPT-Neo 2.7B.

Model	Is it a joke? (Yes %)	Humor Rating (1-5)	Explanation Quality (1-5)
GPT-Neo 1.7B	27.5%	1.68	1.65
GPT-Neo 2.7B	7.5%	1.55	1.75

As Table 5.2 indicates, the performance of the GPT-Neo models in generating humor was inconsistent and somewhat surprising. The GPT-Neo 1.7B model was able to generate outputs that 27.5% of participants recognized as jokes, which is a notable improvement compared to GPT-2. However, when scaling up to the larger GPT-Neo 2.7B model, this percentage dropped sharply to 7.5%. This unexpected decrease might suggest that increasing the model size might not translate to better joke recognition and might have even hindered it.

The humor rating also reflects this inconsistency. The smaller GPT-Neo 1.7B model received an average humor rating of 1.68, slightly better than both GPT-2 models. However, the humor rating for the

GPT-Neo 2.7B model dropped to 1.55, which is lower than the rating for GPT-Neo 1.7B and comparable to the scores of GPT-2. This indicates that the larger model's jokes were perceived as less funny by the participants.

When it comes to the explainability of the jokes, the scores were relatively low for both models. GPT-Neo 1.7B had an explanation quality rating of 1.65, similar to its humor rating, while GPT-Neo 2.7B scored slightly higher at 1.75. This slight improvement in explainability with the larger model suggests that while the jokes generated by GPT-Neo 2.7B were harder to recognize as jokes, they were marginally better at explaining.

Overall, the results for GPT-Neo demonstrate a mixed performance in humor generation. The GPT-Neo 1.7B model outperformed the larger 2.7B model in joke recognition, which challenges the assumption that larger models are always better at such tasks. The decline in both the perceived humor and joke recognition for the larger model maybe indicates that increasing the number of parameters may introduce complexities that do not necessarily contribute to better humor generation.

5.1.2.C T5

Our evaluation continued with the T5 model, which stands for Text-To-Text Transfer Transformer. We used T5 through its HuggingFace library and adjusted the prompts to generate jokes.

The survey results for T5 are summarized in Table 5.3.

Table 5.3: Survey Results for T5.

Model	Is it a joke? (Yes %)	Humor Rating (1-5)	Explanation Quality (1-5)
T5	0%	1.2	2.53

The performance of the T5 model in generating humor was notably poor. None of the jokes generated by T5 were identified as jokes by the participants, resulting in a 0% recognition rate. This indicates that T5 struggled significantly with generating content that was perceived as humorous.

The humor rating for T5 was particularly low, averaging 1.2 on a scale from 1 to 5. This score is the lowest among the models tested, suggesting that the jokes generated by T5 were not only unrecognized as jokes but also considered unfunny by the participants.

On a slightly more positive note, the explanation quality for T5 was somewhat better, with an average score of 2.53. This suggests that while the jokes generated by T5 were not recognized or appreciated as humor, participants found them marginally better at explaining compared to the extremely low explainability scores of the other models. Despite this, the model's inability to produce jokes that were even recognized as such overshadows this minor advantage.

Overall, the T5 model did not perform well in the humor generation task. The lack of joke recognition combined with a low humor rating underscores significant limitations in T5's ability to generate content

that resonates as humorous. This suggests that T5 may not be well-suited for tasks that involve creative content generation, such as humor, and highlights the importance of model selection and fine-tuning for specific applications.

5.1.2.D GPT-3

Our evaluation proceeded with GPT-3, marking our transition to closed-source models. GPT-3 represents a significant leap in model sophistication, and we tested two distinct versions: GPT-3 DaVinci and GPT-3.5-turbo-0125. To access these models, we used the official OpenAI API, which allows us to interact with the models and generate jokes according to our study's needs.

Table 5.4: Survey Results for GPT-3Turbo and GPT-3 DaVinci.

Model	Is it a joke? (Yes %)	Humor Rating (1-5)	Explanation Quality (1-5)
GPT-3Turbo	70%	2.78	2.83
GPT-3 DaVinci	7.5%	1.55	1.5

Table 5.4 reveals a marked contrast between the two GPT-3 versions. GPT-3 Turbo demonstrated a significantly better performance in humor generation. With 70% of the jokes being recognized as such, it outperformed all previously tested models by a considerable margin. This indicates that GPT-3 Turbo was notably more effective at producing content that resonated as humor.

The humor rating for GPT-3 Turbo was 2.78 on a scale from 1 to 5, which, while still modest, is a notable improvement compared to the other models. This suggests that the jokes generated by GPT-3 Turbo were perceived as somewhat funny by participants, reflecting a higher effectiveness in humor generation.

In terms of explanation quality, GPT-3 Turbo received an average score of 2.83. This indicates that while the humor was more recognizable, there is still room for improvement in making the generated content both funny and more explainable.

On the other hand, GPT-3 DaVinci's results were less favorable. Only 7.5% of its outputs were identified as jokes, reflecting a significant challenge in generating humor. The humor rating for GPT-3 DaVinci was 1.55, and the explanation quality was slightly lower at 1.5. These results suggest that the DaVinci version struggled both with creating humor that resonated with participants and with producing content that easily explained why the article had that particular label.

In summary, while GPT-3 Turbo showed a strong ability to generate content recognized as jokes and achieved better humor ratings, GPT-3 DaVinci lagged behind significantly. This highlights the effectiveness of the GPT-3 Turbo model in humor generation tasks, demonstrating that it can produce more engaging and recognizable humorous content compared to its counterpart.

5.1.2.E GPT-4

The final model in our study is GPT-4, the latest advancement from OpenAI. To use GPT-4, we accessed it via the official OpenAI API, similar to our approach with GPT-3. GPT-4 represents the cutting edge of language models, incorporating significant improvements in comprehension, generation, and overall performance.

The survey results for GPT-4 are summarized in Table 5.5.

Table 5.5: Survey Results for GPT-4.

Model	Is it a joke? (Yes %)	Humor Rating (1-5)	Explanation Quality (1-5)
GPT-4	75%	2.83	3.93

Table 5.5 highlights GPT-4's impressive performance in humor generation. With 75% of the jokes being recognized as such, GPT-4 outperformed all previous models tested in this study. This high recognition rate indicates that GPT-4 is highly effective at producing content that participants perceive as humorous.

The humor rating for GPT-4 averaged 2.83 on a scale from 1 to 5. While this score remains in the lower end of the scale, it represents a notable improvement over previous models, reflecting that the jokes generated by GPT-4 were considered somewhat funny by participants.

Remarkably, GPT-4 excelled in explanation quality, achieving an average score of 3.93. This score is the highest among all models tested, suggesting that GPT-4's jokes were not only recognized as humor but also easier to understand the explanation they conveyed. This high explanation quality indicates that GPT-4 effectively balances humor with clarity.

Overall, GPT-4 demonstrated substantial advancements in humor generation compared to earlier models. Its high joke recognition rate, combined with improved humor and explanation scores, underscores its capability to produce engaging and understandable humorous content. This marks GPT-4 as a leading model in the field of language generation, particularly for tasks that involve nuanced and effective humor creation.

5.1.2.F Conclusion

After a comprehensive evaluation of various language models for humor generation, we have selected GPT-4 as the model for our system. The decision was based on several key factors highlighted in our study.

GPT-4 demonstrated the highest performance across all tested models, with 75% of its generated jokes recognized as such by participants, which is the best result among the models evaluated. This strong performance in joke recognition indicates that GPT-4 excels at producing content that resonates

as humor.

Furthermore, GPT-4 achieved a humor rating of 2.83 and an explanation quality score of 3.93, both of which are the highest in our study. These results make GPT-4 the most suitable choice for our system, given its demonstrated effectiveness in humor generation and clarity. The model's ability to balance humor with comprehensibility ensures that our system will provide engaging and intelligible content. Thus, GPT-4's advanced capabilities and performance metrics justify its selection as the core component of our humor generation system.

5.2 Joke generation Approaches

Based on the comparison study, we selected GPT-4 for our joke generation approaches. Since OpenAI's models are proprietary and not open-source, we used their API to integrate the chosen model into our project.

5.2.1 Prompt Chaining

Our initial attempt at generating humor using the selected model involved implementing the humor algorithm proposed by Joe Toplyn [52], as discussed in Section 3.1.2. To adapt the algorithm to our context, we tested various prompts to determine the best possible outcomes. Specifically, we evaluated whether it was more effective to include both the claim and the explanation in the first step of the algorithm or to use only the claim. Additionally, we also considered adding an extra step to the algorithm to modify the jokes, ensuring they remain contextually relevant to the news articles and retain their humor.

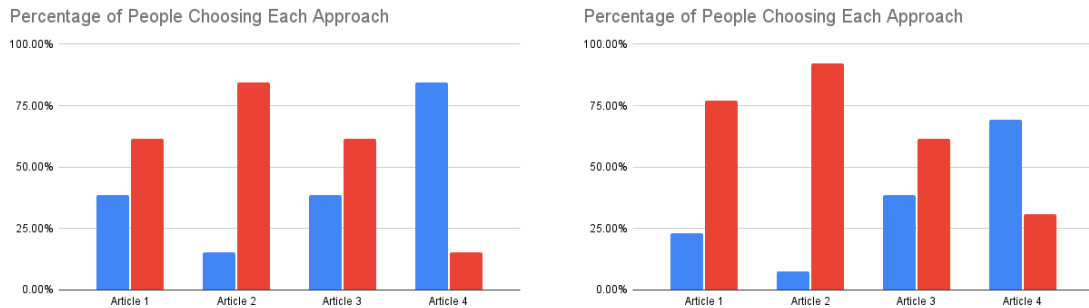
To identify the best prompt and minimize potential bias, we conducted a study with thirteen participants. The group comprised four females and nine males, with an average age of 21. Participants' education levels ranged from high school to a master's degree, with all either holding or pursuing degrees in engineering. Each participant reviewed two outputs from different prompts across four distinct news articles. The results are as follows:

As shown in Figure 5.1(a), for 3 out of the 4 articles, participants preferred the prompt that included both the claim and the explanation in the first step of the algorithm. Therefore, we adopted this approach. Similarly, Figure 5.1(b) indicates that the modified algorithm with the additional step was preferred for 3 out of the 4 jokes. Consequently, we incorporated this extra step into the algorithm.

So our final algorithm had these prompts:

1. Identify Key Nouns:

- Choose the two most conspicuous nouns or noun phrases in the given text. Exclude any personal names, numbers, years, and dates.



(a) Testing results: claim only (blue) vs. claim plus explanation (red).

(b) Testing results: original algorithm (blue) vs. algorithm with extra step (red).

Figure 5.1: Evaluation of prompts and algorithm modifications.

- **Prompt:** "Choose the two most conspicuous nouns or noun phrases in the following text, excluding any person names, numbers, years, and dates: Claim + Explanation"

2. Generate Associations:

- Generate a list of associations or related words for each of the chosen nouns.
- **Prompt:** "Generate a list of associations for each of these words: Chosen Nouns"

3. Form Punchline:

- Combine one association from each list to form potential punchline.
- **Prompt:** "Combine one association from each list to form a punchline: Association List"

4. Generate Joke Candidates:

- Create a joke candidate by considering the veracity of the label. The joke should be based on the topic and one of the punchline candidates.
- **Prompt:** "This news article has a label of {Label}. Generate a joke candidate, taking into account the veracity of the label, based on the topic and one of the punchline candidates. Topic: {Claim} + {Explanation} Punchlines: {Punchlines}"

5. Refine and Finalize the Joke:

- Make small changes to the joke if necessary to ensure it explains why the topic has the specified claim. Only output the final joke.
- **Prompt:** "Make small changes, if necessary, to the joke so it explains why the topic has this claim. Only output the final joke. Topic: {Claim} + {Explanation}; Label: {Label}; Joke: {Joke}"

To illustrate the effectiveness of this approach, here is an example of a joke generated using our final algorithm, based on an article labeled as “False” regarding a claim about the Las Vegas Sphere.

Example:

- **Claim:** The Las Vegas Sphere displayed an Israeli flag.;
- **Label:** False;
- **Explanation:** This image was altered, the venue said.;

Using the refined algorithm, the following steps led to the creation of a contextually relevant and humorous joke:

1. **Identify Key Nouns:** Las Vegas Sphere and Israeli Flag;
2. **Generate Associations:**
 - Las Vegas Sphere: Entertainment, Arena, Concert Venue, Las Vegas, etc.;
 - Israeli Flag: Blue and white, Star of David, National emblem, Jerusalem, etc.
3. **Form Punchline:** “The Las Vegas Sphere, with its capacity of 17,500 seats and LED exoskeleton, is setting to become an entertainment attraction as distinct and pride-inducing as the blue and white colors of the Israeli flag are to its citizens”;
4. **Generate Joke Candidate:** “Remember when the Las Vegas Sphere displayed an Israeli flag? Yeah, that was as real as the chance of actually leaving Vegas with more dollars than you came with!”;
5. **Refine and Finalize the Joke:** “Do you recall seeing the Las Vegas Sphere displaying an Israeli flag? Yeah, that image was as altered as the dream of actually leaving Vegas with more dollars than you came with!”.

This approach, as illustrated, consistently helped generate jokes that were not only humorous but also aligned with the context and the article’s label.

5.2.2 First Fine-tuning Approach

Our second approach aimed to leverage the high performance observed in our comparison study by fine-tuning the top-performing model, GPT-4. However, due to access restrictions and the high-tier usage requirements for fine-tuning GPT-4, we were unable to proceed with this model. Given these constraints, we turned our attention to the next best option from our study: GPT-3.5-turbo-0125 which also exhibited excellent performance during our initial evaluations, making it an ideal candidate for our fine-tuning efforts.

5.2.2.A Dataset

We used the rJokes dataset ¹ built by Weller et al. [66] to fine-tune the model. This dataset is built from scraping the subreddit of r/Jokes, an online forum. It includes valuable information extracted from Reddit's servers, including the creation timestamp, post titles, the content of the jokes, and the number of upvotes received – a metric regarded as a reliable indicator of the joke's reception within the community. The rJokes dataset comprises approximately 550,000 entries, offering a diverse array of jokes that have resonated with the online audience. However, as fine-tuning the model incurs costs, we optimized our dataset by ordering the jokes based on their upvotes and selecting the top 100,000 jokes for the fine-tuning process. This subset still provides a substantial amount of high-quality jokes for the model to learn from.

With this curated selection of jokes, we proceeded to the data pre-processing stage to ensure optimal input quality for the model:

- **Length Filtering:** We removed all jokes that exceeded a length of 1300 characters. This decision was made to avoid lengthy textual explanations, which are typical of traditional fact-checking websites. Given the context of jokes, brevity is crucial as longer jokes might deter people from reading the entire content;
- **Redundancy Check:** Duplicate entries and highly similar jokes were identified and removed to prevent the model from overfitting to specific joke formats;
- **Offensiveness Filtering:** To ensure the dataset was free from offensive content, we implemented a two-step filtering process. First, using a list of Google's profanity words², we designed a script that scanned all the jokes and removed any that contained words from this list. Second, we used a Python library to further review the jokes and filter out any that were deemed inappropriate.

After processing the data, our dataset contained 63,000 jokes. We conducted a comprehensive dataset analysis and visualization with the final version, examining various aspects including the average length of the entries, the longest and shortest entries, and creating a word cloud of all the entries, seen in Image 5.2. Our findings are summarized below:

- **Average Length:** 189.195 characters;
- **Longest Joke:** 1290 characters;
- **Shortest Joke:** 60 characters.

¹<https://github.com/orionw/rJokesData>.

²<https://github.com/coffee-and-fun/google-profanity-words>

we also analyzed various prompts to identify which produced the best results. Initially, the prompt used during fine-tuning was “Write me a joke for this topic,” followed by presenting the joke to the model. Our study aimed to evaluate whether it was better to use this single prompt and pass the claim as the topic, or to use an additional prompt asking the model to ensure the joke is funny and contextually relevant to the article. The results of this comparison are illustrated below, in Figure 5.3:

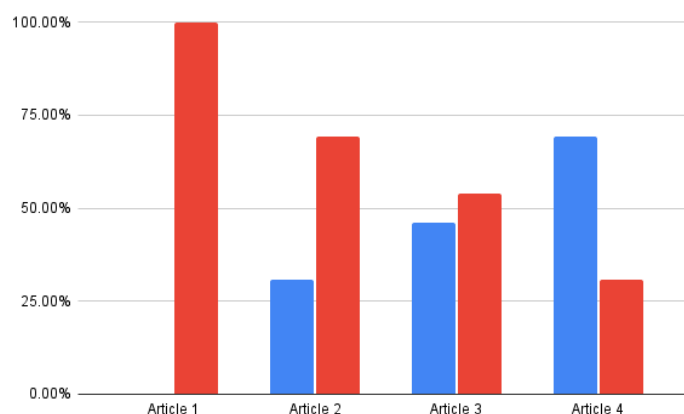


Figure 5.3: Percentage of Respondents Choosing Each Approach Results: Only one prompt (blue) vs. Two prompts (red).

As shown in Figure 5.3, the jokes generated using two prompts outperformed those generated using a single prompt in 3 out of 4 news articles. Therefore, we decided to adopt the two-prompt approach to enhance the quality and relevance of the jokes.

5.2.2.D Hyperparameter Optimization

After refining the prompts, we shifted our focus to optimizing the hyperparameters to further enhance the quality of the generated content. Our goal was to determine the most effective configuration of hyperparameters, specifically by altering the temperature and top-p settings, as these have the most significant impact on the output. We reached this conclusion after consulting the OpenAI API reference [68], which provides detailed guidance on the available hyperparameters and their respective effects on model behavior.

We began by conducting a systematic evaluation of the temperature parameter. We selected three initial temperature values—0.2, 0.6, and 0.9—which provided a broad range of sampling variability. To assess the impact of each value, we generated jokes for fifteen different news articles, varying only the temperature setting. During this process, we observed that a temperature of 0.9 frequently led to hallucinations in the generated content. To address this issue, we included an additional temperature setting of 0.75 in our tests. After generating jokes at each of the four temperature levels, we enlisted the assistance of two independent evaluators. They reviewed the jokes and selected the best one for each

article. The temperature that produced the most frequently selected jokes was deemed optimal.

With the optimal temperature established, we proceeded to optimize the top-p parameter using a similar methodology. We tested values of 0.2, 0.6, and 0.9, while holding the temperature constant at the previously determined optimal value. Unlike with temperature, we did not observe significant discrepancies between top-p values of 0.6 and 0.9, so we continued testing with these settings.

Ultimately, the hyperparameters that produced the highest quality jokes were determined to be:

- **Temperature** – 0.6;
- **Top-p** – 0.2.

After concluding the hyperparameter optimization, we can illustrate the effectiveness of this approach by presenting a sample joke generated by it. Below processed through our two-prompt strategy and with the optimal hyperparameters is a joke for a claim about the Covid-19 vaccine.

Given the claim:

Claim: “All these athletes are dropping dead on the field after receiving the COVID-19 vaccination.”

The model generates the following joke:

Joke: “I’m not sure why they’re getting vaccinated on the field, though.”

In this example, the model uses irony to highlight the absurdity of the claim, creating humor by imagining an exaggerated scenario where vaccinations take place right on the field. This contextually relevant and lighthearted response showcases this approach’s ability to inject humor effectively while addressing the claim.

5.2.2.E Domain-Specific Fine-Tuning

After conducting the initial fine-tuning with the entire dataset, we raise the hypothesis that using a more focused subset of jokes—specifically those related to a particular topic such as war or COVID-19—might enhance the model’s ability to explain news articles on those specific topics. To test this hypothesis, we created a smaller dataset containing 50 jokes exclusively about COVID-19 and used it for fine-tuning the model. The outcomes of this domain-specific fine-tuning are discussed in detail in Section 6.2.3

5.2.3 Second Fine-tuning Approach

Our final approach involves fine-tuning the GPT-3.5-turbo-0125 model, this time using a dataset specifically curated for the task of generating jokes based on a claim, label, and explanation.

5.2.3.A Dataset

To train our final system to generate jokes based on a claim, a label, and an explanation, we created a custom dataset incorporating these components. This approach aims to help the model better understand the expected behavior and produce more relevant and humorous content.

Initially, we needed to gather a substantial collection of news articles that included the necessary elements: a claim, a label, and an explanation. Referring to Section 3.2, we leveraged a master's thesis that had previously explored creative explanations, such as poems [61]. This thesis provided a dataset of 1000 news articles. We decided to expand upon this existing dataset by including more recent news articles, thereby enhancing its relevance and diversity.

To collect these additional news articles, we developed a web-scraping script targeting the Politifact website. The script navigates through the site and extracts articles containing the three critical components required for our dataset: the claim, the label, and the "If your time is short" summary. By augmenting the initial dataset with these additional entries, we increased the total number of articles to 1500. Given the complexity of humor generation, we do not intend to manually create jokes for all 1500 entries. However, having a larger pool of news articles provides more material from which to craft jokes, thereby increasing the likelihood of producing high-quality humorous content.

As this is no easy task and since we do not want to introduce any type of bias towards the jokes made, we opted to seek assistance in constructing this dataset. We formed a group of six individuals, all male, aged between 22 and 25. Among them, two held Bologna Master degrees, while the remaining four had Bologna Bachelor degrees, all in computer science. To facilitate the collaboration, we prepared a Google Sheets document containing the news article dataset and added two new columns: "Joke" and "Author". We then shared the link to this dataset with everyone involved in creating jokes.

With this collaborative effort, we managed to write 100 jokes. These 100 jokes had the following label distribution: 48.9% False, 33.3% Pants on Fire, 7.8% True, 4.4% Half True, 3.3% Barely True, 2.2% Mostly True. As observed, 82.2% of the jokes were made for articles classified as misinformation. Despite this being a small subset of the entire dataset, it reflects a significant focus on misinformation, which is consistent with the nature of the broader dataset, where 62.1% of the articles are classified as misinformation.

In Table 5.6, we present the number of jokes each person wrote and their education degree. As you can see, the total number of jokes presented in the table is not 100, but 63. This discrepancy is due to the fact that we also searched online for jokes already made to the articles in our dataset. If we found suitable jokes, we included them. Using this method, we managed to gather 37 additional jokes for our dataset. This dual approach of creating original jokes and sourcing existing ones ensured a rich and diverse collection of humorous content, enhancing the dataset's overall quality and variety. This meticulous approach to dataset creation helps ensure that the model receives diverse and representative

examples, thereby improving its ability to generate humorous content while minimizing bias.

Person	Number of Jokes Written	Education Degree
Person 1	4	Bologna Bachelor
Person 2	4	Bologna Bachelor
Person 3	15	Bologna Bachelor
Person 4	3	Bologna Master
Person 5	30	Bologna Bachelor
Person 6	7	Bologna Master

Table 5.6: Number of jokes written by each person and their education degree.

5.2.3.B Fine-tuning

With our dataset complete, we could now start the fine-tuning process. We needed to convert our dataset to the same format as in the previous section, but this time with a different prompt. We tested five different prompts to determine which would produce the best results. We asked three people for their opinions on the jokes generated by the different prompts, and their responses were unanimous. The best prompt used the “System description” feature and it consisted of the following:

- “You are a system designed to generate jokes that explain why a news article has a certain label. Your jokes should be funny, insightful, and directly related to the given claim, label, and article explanation.”

The actual prompt was:

- “Claim: claim Label: label Article Explanation: explanation Write a humorous joke that explains why this article has received its label.”

With the prompt and dataset ready, we used 90 out of the 100 entries in the dataset for training. The remaining 10 entries were reserved for a final questionnaire to assess whether the jokes were good enough compared to human-made ones and if the jokes generated by the system could pass as human-made.

5.2.3.C Hyperparameter Optimization

To ensure optimal performance for this approach, we conducted a comprehensive hyperparameter study, aimed at selecting the most effective settings. This process followed the same systematic methodology outlined in Section 5.2.2.D, where we rigorously tested different combinations of hyperparameters to

determine the configuration that would yield the highest quality results. After extensive testing and evaluation, we identified the settings that consistently produced the best outcomes:

- **Temperature** – 0.75;
- **Top-p** – 0.9.

By fine-tuning these hyperparameters, we were able to significantly improve the performance of the model, leading to the generation of high-quality jokes that aligned well with the intended tone and context of the task.

With the optimal prompt and hyperparameters established, we can demonstrate the effectiveness of this approach by showcasing an example joke generated from a news claim, label, and article explanation. Below, we provide a headline, label, and article explanation used as input, followed by the model's humorous response.

- **Claim:** "No live audience at CNN debate means we are watching a scripted movie.";
- **Label:** False ;
- **Article Explanation:** President Joe Biden and former President Donald Trump agreed not to have a live studio audience during their June 27 debate in Atlanta. Both candidates spoke about their debate performances in campaign rallies the day after.

With these inputs, the model generated the following joke:

Joke: "The debate was not scripted, but i think the two candidates should have rehearsed a little bit more."

In this example, the approach successfully integrates humor by suggesting that while the debate wasn't scripted, perhaps the candidates' performances could have benefited from a bit more rehearsal. This playfully hints at the sometimes chaotic nature of political debates, adding a layer of irony to the claim without reinforcing misinformation. The result demonstrates the ability to craft contextually relevant humor that is both insightful and amusing, effectively responding to the claim in a way that engages the audience.

5.3 Moral Classifier

In this section, we address the crucial aspect of ensuring that the jokes generated by our system are morally acceptable. Given the sensitive nature of explaining and commenting on news articles, it is crucial that the generated humor adheres to ethical standards. Morally correct jokes are a must for

this thesis, as we cannot afford to produce explanations that are offensive, inappropriate, or ethically questionable. Therefore, we incorporated a moral classifier to screen the jokes, ensuring they are not only funny and insightful but also align with accepted moral and ethical norms. This approach guarantees that our system generates content that is both entertaining and responsible.

5.3.1 Moral Foundations

Before delving into the specifics of our moral classifier, it is important to discuss the moral foundations theory that forms the basis of the classifier. This theory, as mentioned in Section 3.1.1, identifies five core moral factors or foundations found across cultures [43]. These foundations are:

- Care/Harm;
- Fairness/Cheating;
- Loyalty/Betrayal;
- Authority/Subversion;
- Purity/Degradation.

Understanding these foundations allows us to create a moral classifier that evaluates the jokes generated by our system, ensuring they align with widely accepted ethical standards. The classifier uses these moral dimensions to assess whether a joke is morally acceptable, thus preventing the generation of content that could be harmful, unfair, disloyal, disrespectful of authority, or impure. By grounding our approach in moral foundations theory, we aim to produce jokes that are not only humorous and insightful but also ethically responsible.

5.3.2 Dataset

With these constraints set, we need to find a dataset to train our classifier. We decided to use the Moral Foundations Twitter Corpus as described by Hoover et al. [42]. This dataset contains over 35,000 entries, each corresponding to a tweet ID collected from the social media platform “X” (formerly known as Twitter). Each tweet in the dataset has been annotated for moral foundations, providing a rich resource for understanding moral expressions in social media.

The annotations in the dataset cover 11 categories: 10 categories correspond to the various moral foundations identified by moral foundation theory, and an additional category labeled “Non-moral” is used for tweets that do not express any of the identified moral foundations. This extra category is particularly useful as it helps differentiate between moral and non-moral content, offering a clearer understanding of how moral language is used on social media.

Each tweet in the dataset was annotated by at least three annotators. The final annotation for each tweet was determined by the majority vote among the annotators, ensuring the reliability and accuracy of the annotations. This multi-annotator approach reduces individual bias and increases the robustness of the dataset.

Since the dataset includes tweet IDs rather than the full text of the tweets, in compliance with X's data sharing policies, we needed to use the official X API to retrieve the full tweet text. To accomplish this, we first familiarized ourselves with the API by reviewing the documentation provided by the "X Developer Platform" [69]. Based on this, we developed a custom script to fetch the text of the tweets using their corresponding IDs. However, due to the cost associated with API access and the rate limits imposed by the platform, we were only able to retrieve the full text for 10,000 tweets. Given these constraints, it was essential to carefully select a balanced subset of the dataset. This selection process was critical to ensure that our classifier could learn effectively from the data, covering a representative range of moral and non-moral categories.

After selecting the subset, we ran our script to fetch the tweet texts, which formed the final dataset we would use for training and evaluation. This balanced and carefully curated dataset provided a solid foundation for our classifier, enabling it to perform well despite the limitations in the number of tweets we could retrieve.

5.3.3 Training and evaluation

With the dataset ready, we started developing the code to train our moral classifier. We decided to use BERT for this task due to its proven effectiveness in natural language understanding. The training process involved several key steps to ensure that the model could accurately classify tweets according to the moral foundations.

We began by splitting the dataset into training and testing sets, ensuring that our classifier would be trained on one subset of the data and evaluated on another to measure its performance. We used an 90-10 split to create the training and testing sets, respectively. Then, each tweet was tokenized using the BertTokenizer. The tokenizer converts the tweets into a format suitable for BERT, including adding special tokens, truncating or padding to a fixed maximum length (256 tokens), and generating attention masks and token type IDs.

We took the BERT model and added a simple classification layer on top, which helps turn BERT's output into predictions for the 11 labels (10 moral foundations and one non-moral category). BERT's processed text is passed through this layer to generate the final predictions.

We trained the model for 3 epochs, using Binary Cross-Entropy as the loss function, which is suitable for multi-label classification tasks. During each epoch, the model's parameters were updated using the Adam optimizer, with a learning rate of $1e-5$. To evaluate the model's performance, we used several met-

rics, including accuracy, Hamming loss—which measures the proportion of misclassified labels across all instances—and a confusion matrix. These metrics provided a comprehensive understanding of how well the model was able to classify the tweets into the different moral foundation categories. These metrics are visible in Image 5.4(a)

After training the initial model, we opted for an ensemble approach by incorporating a second model, RoBERTa, trained on the same task. The final output is obtained by averaging the results from both models. We chose this strategy because ensemble methods typically offer greater generalization power compared to individual models [70]. By ensembling we leverage the strengths of different models, reducing the likelihood of overfitting by averaging out the noise and errors from individual models. This approach often results in more robust predictions, as the ensemble tends to capture a broader spectrum of patterns in the data. Additionally, ensembling can mitigate the biases inherent in individual models, leading to improved accuracy and stability in the final predictions. By combining models like BERT and RoBERTa, which may capture different aspects of the data, the ensemble can provide a more comprehensive understanding, thus enhancing performance. We trained the second model under the same conditions as the first, and its performance metrics are shown in Image 5.4(b).

```
Validation Loss: 0.2846923344768584
Accuracy: 0.3480392156862745
Hamming Loss: 0.117825311942959
Classification Report:
```

	precision	recall	f1-score	support
0	0.70	0.61	0.65	196
1	0.73	0.58	0.65	228
2	0.89	0.77	0.83	746
3	0.69	0.72	0.70	165
4	0.81	0.49	0.61	119
5	0.80	0.54	0.64	237
6	0.88	0.52	0.65	118
7	0.64	0.61	0.63	150
8	0.48	0.52	0.50	82
9	0.81	0.65	0.72	266
10	0.81	0.73	0.77	226
micro avg	0.79	0.66	0.72	2533
macro avg	0.75	0.61	0.67	2533
weighted avg	0.79	0.66	0.71	2533
samples avg	0.84	0.73	0.75	2533

(a) BERT Classifier Metrics.

```
Validation Loss: 0.28495225850492717
Accuracy: 0.3392156862745098
Hamming Loss: 0.12130124777183601
Classification Report:
```

	precision	recall	f1-score	support
0	0.59	0.62	0.60	196
1	0.67	0.54	0.60	228
2	0.84	0.86	0.85	746
3	0.70	0.64	0.66	165
4	0.85	0.54	0.66	119
5	0.91	0.45	0.60	237
6	0.91	0.43	0.59	118
7	0.72	0.51	0.60	150
8	0.56	0.45	0.50	82
9	0.74	0.78	0.76	266
10	0.72	0.79	0.76	226
micro avg	0.76	0.68	0.72	2533
macro avg	0.75	0.60	0.65	2533
weighted avg	0.77	0.68	0.71	2533
samples avg	0.81	0.75	0.75	2533

(b) RoBERTa Classifier Metrics.

Figure 5.4: Moral Classifier Metrics.

After training the models, we designed the following structure for the moral classifier, as illustrated in Image 5.5. The moral classifier takes jokes as input and uses each model to predict the probability of each moral foundation being present in the joke. These probabilities are then averaged to obtain the final scores for each moral foundation.

The classifier includes a two-step evaluation process. First, a moral check is performed: if the probability for either the Harm or Degradation moral foundations exceeds 0.8, the joke is flagged as morally unacceptable. We focus on these two foundations because they directly relate to actions or statements

that can cause emotional harm or offense. While other moral foundations can also lead to harm, the potential for significant emotional damage is more acute with Harm and Degradation. Thus, concentrating on these foundations allows us to address the most immediate and severe forms of emotional harm.

Following the moral check, a non-moral check is conducted. This step verifies whether the probability for a non-moral foundation exceeds 0.8. If this threshold is met, the joke is deemed acceptable, regardless of the moral check outcome. However, if the non-moral probability is below 0.8 and the joke triggers a high probability for Harm or Degradation, the joke is filtered out. This moral classifier approach prioritizes detecting content with the most significant risk of causing emotional harm while allowing exceptions when a non-moral foundation is strongly represented. This dual-check system ensures a balanced approach, filtering out potentially harmful jokes while preserving those that may still be acceptable based on their broader context.

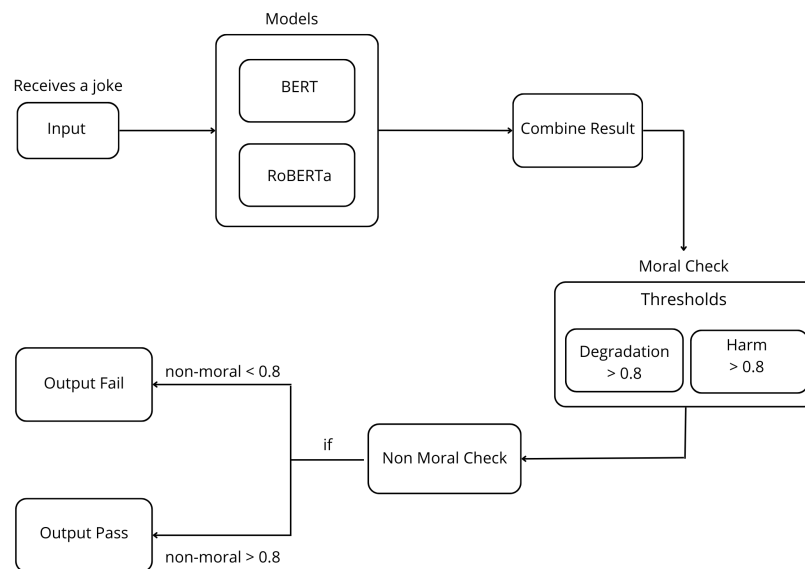


Figure 5.5: Moral Classifier Architecture.

5.3.4 Example of Morally Inappropriate Joke Detection

With our moral classifier fully implemented, we can illustrate its effectiveness by presenting an example of a joke that it flagged as morally inappropriate. This example demonstrates how the classifier's dual-check system helps screen jokes that may contain elements of harm or degradation, ensuring that the generated content aligns with ethical standards.

Given the input:

- **Claim:** "Jews did 9/11"
- **Label:** Pants on Fire

- **Article Explanation:** Al-Qaida terrorists were responsible for the Sept. 11, 2001, attacks, not ‘Jews’ or Israel.

The model generates the following joke:

Joke: “Al-Qaida terrorists were responsible for the Sept. 11, 2001, attacks, not ‘Jews’ or Israel. But you gotta admit, these were some explosive accusations.”

In this instance, the classifier identified a high probability of a Harm violation, flagging the joke as morally inappropriate. The humor attempt relies on wordplay about an event that is both deeply sensitive and traumatic. By referencing “explosive accusations,” the joke risks trivializing the gravity of the 9/11 attacks, potentially causing distress to those affected by the tragedy.

This example underscores the classifier’s role in upholding ethical standards, screening jokes for sensitive or potentially distressing content. In doing so, it ensures that humor is presented in a way that is respectful, minimizing the risk of harm while still striving to provide entertainment responsibly.

5.4 Joke Classifier

In this section, we explore the development of our joke classifier, detailing the datasets employed for training and the various approaches we analyzed. We cover the rationale behind our model choices, the implementation, and the evaluation criteria used to ensure we achieved the most effective and accurate classifier possible.

5.4.1 Classifier Options

We now delve into the various classifier options that were considered and thoroughly analyzed during our research. We will provide a comprehensive overview of each method, discussing the specific implementation strategies employed, and any challenges or advantages encountered along the way. In Section 5.4.2, we analyse the performance of each classifier.

5.4.1.A Fine-tuned BERT Classifier

Our initial approach involved fine-tuning a pre-trained model, such as RoBERTa or BERT on, a large dataset containing 1.7 million rated jokes. This dataset, known as the Jester dataset [71], features continuous ratings ranging from -10 to 10, with 150 different jokes rated by 59,132 users. We hypothesized that with such a vast number of ratings, the model could effectively learn to evaluate humor.

To begin, we needed to prepare the dataset for training. The original dataset was provided in two separate CSV files: one containing the user ID, joke ID, and the rating given by the user, and the other

mapping the joke IDs to the actual text of the jokes. Our task was to merge these files to create a new CSV file that contained each joke's text along with the corresponding ratings. This consolidated dataset allowed us to pair each joke with its multiple ratings, enabling the model to learn from a wide range of humor assessments.

With the dataset prepared, we initiated the training of a BERT model. To optimize the model's performance, we incorporated a dropout layer, which helps to prevent overfitting by randomly disabling a fraction of neurons during training. Additionally, we applied a ReLU (Rectified Linear Unit) activation function to introduce non-linearity, which enhances the model's ability to learn complex patterns. Given the large size of the dataset, we set a learning rate of $1e - 5$. This small learning rate ensures that the model's weights are updated incrementally, reducing the risk of overshooting the optimal values and improving the stability of the training process.

Considering the extensive size of the dataset, training the model on the entire dataset at once would have been computationally prohibitive. Therefore, we opted to train the model one epoch at a time. After each epoch, we saved the model's weights, loss, and all other relevant parameters. This approach allowed us to incrementally build the model's understanding of the data, while ensuring that we could resume training from the last checkpoint if needed, without exhausting our computational resources.

5.4.1.B GPT-4 Powered Classifier

The second option we explored involved the system described in Section 3.3, which utilized GPT-4 to evaluate jokes [64]. In this implementation, we made several adjustments to the original system configuration and the examples used. Specifically, we refined the system descriptions to better align with our classification objectives, incorporating a broader range of humor types that the system could identify as funny.

Additionally, the examples provided in the prompts were updated to more accurately reflect the nuances of the joke types being classified. These modifications aimed to enhance the system's sensitivity to different humor styles, thereby improving the overall accuracy of the joke classification task.

Despite these changes, we also tested the original prompts of the system to compare their results with our modified approach.

5.4.1.C SSTH based Classifier

The final option we explored was based on the system also mentioned in section 3.3. The original system was trained to predict whether a given text was humorous or not. However, the dataset they used for training also contained ratings associated with each joke. With this in mind, we adapted their code, available on GitHub⁴, so the system could rate jokes instead of merely determining whether a

⁴<https://github.com/Moradnejad/ColBERT-Using-BERT-Sentence-Embedding-for-Humor-Detection>

sentence is humorous.

This involved altering the training data fed into the model. Instead of binary labels, the model was trained on joke-rating pairs from the dataset. We did not make any changes to the model architecture or the underlying code, but by simply providing the model with continuous rating data, we enabled it to output a rating for each joke rather than a binary classification.

This approach leveraged the existing model's ability to understand humor while extending its functionality to generate more nuanced assessments in the form of ratings. While the modification was relatively simple, it allowed us to repurpose the system for our specific needs without extensive re-engineering.

5.4.2 Performance Comparison

With all the systems implemented, we were prepared to evaluate and compare their performance. While a common approach would involve assessing the accuracy of each system's ratings by directly comparing them to the ratings provided in the dataset, our objectives required a different evaluation metric.

In our specific application, the primary function of the joke classifier is not merely to assign accurate ratings, but to consistently identify the funniest option when presented with three jokes. This ranking ability is crucial, as the classifier's effectiveness hinges on its capacity to mirror human judgment in determining which joke is more humorous in a given set. Therefore, instead of focusing solely on rating accuracy, we prioritized the system's ability to rank jokes correctly according to perceived funniness.

To evaluate this capability, we designed a test in which the system was presented with pairs of jokes. Its task was to select the funnier joke from each pair. We then compared the system's choices with the selections made by a panel of five human annotators, who also chose the funnier joke from the same pairs. This comparison allowed us to measure how well the system's judgments aligned with those of human evaluators.

By analyzing the agreement between the system's choices and the majority choice of the human annotators, we were able to assess the effectiveness of each implementation. This method provided a clear indication of how well the system performs in practical scenarios where humor is subjective and context-dependent. Through this comparative analysis, we identified which system configuration most closely approximates human humor judgment, guiding our selection of the most effective model for our needs.

To evaluate our system, we created a spreadsheet containing 30 pairs of entries for the annotators to review. We randomly selected 54 jokes from the Reddit dataset used in Section 5.2.2 and included an additional 6 entries consisting of news article headlines. These non-humorous headlines were added to ensure that the system can accurately differentiate between a joke and a regular sentence.

Once all entries were annotated, we used the joke classification options provided by the system. After collecting the choices, we calculated the Cohen Kappa coefficient to measure the agreement between

each annotator and the system, as well as between the system and the average of the annotators' choices.

The Cohen Kappa coefficient is a statistical measure of inter-rater agreement for qualitative items, taking into account the possibility of agreement occurring by chance. Kappa values range from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond what would be expected by chance, and negative values suggest disagreement. Generally, a Kappa value above 0.6 is considered substantial, while values above 0.8 indicate near-perfect agreement.

With all entries annotated and our evaluation metric selected, we proceeded to evaluate each system. The results are presented in Table 5.7.

Table 5.7: Cohen Kappa Coefficients between Systems and Annotators.

System	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Average
Option 1	0.19	0.03	-0.22	-0.06	-0.06	0.01
Option 2 (Original)	0.45	0.17	0.42	0.27	0.45	0.38
Option 2 (Modified)	0.52	0.23	0.36	0.45	0.39	0.46
Option 3	0.51	0.65	0.35	0.63	0.51	0.59

The results reveal distinct patterns in the performance of the different systems. The first system shows the lowest performance, with Kappa values that are either close to zero or negative. The average Kappa for this option is 0.01, indicating almost no agreement with the annotators. The negative Kappa values suggest that the system not only failed to agree with the annotators but also actively disagreed. This poor performance could be attributed to the presence of multiple ratings for the same joke and that might have introduced noise, confusing the model rather than enhancing its learning capability.

The original version of the second system shows moderate improvement over the first option, with an average Kappa of 0.38. While this is a significant improvement, the Kappa values still indicate only fair to moderate agreement with the annotators. The modified version of the second system further improves upon the original, achieving an average Kappa of 0.46. The modifications made to this system appear to have enhanced its ability to align with annotator judgments. This improvement suggests that the modifications allowed the system to give better ratings across different types of jokes, leading to more consistent performance. However, the Kappa values still fall short of substantial agreement, indicating room for further enhancement.

The third system stands out as the best performer, with an average Kappa of 0.59. This value indicates a substantial level of agreement with the annotators. The relatively high Kappa values across all annotators suggest that this system is more robust and consistent in rating humor, making it less susceptible to individual biases.

The analysis of these results highlights the challenges of developing an accurate joke classification

system. The first option proved to be ineffective, likely due to issues in the training data. The modifications made to the second option provided a noticeable improvement, but the third option was clearly the most effective, achieving the highest level of agreement with the annotators. Given these results, we selected the third option as the final joke classifier for our system, as it consistently provided the most reliable performance in capturing the nuances of humor.

To demonstrate the classifier's effectiveness, consider the following three jokes generated by our humor generation techniques, which were evaluated by both our classifier and five human annotators:

1. **Joke 1:** "Did you hear the Supreme Court canceled universal vaccination in the US? Yeah right, and Bigfoot is my tennis partner! They say the Supreme Court building is majestic, but it can't seem to keep up with the pace of the WHO's vaccine mandates!";
2. **Joke 2:** "Why did the Supreme Court justices bring a gavel to their COVID-19 vaccination? Because they wanted to rule out any false news about them canceling universal vaccination!";
3. **Joke 3:** "If the US Supreme Court had actually canceled universal vaccination, the news would have been all over social media. After all, who needs real news outlets when you have angry uncles on Facebook?".

When asked to choose the funniest joke, the classifier selected Joke 3. Four out of five human annotators agreed, with the fifth selecting Joke 2. This exercise highlights the classifier's ability to align with human humor preferences, demonstrating its capability to identify and prioritize nuanced comedic elements.

5.5 System Integration and Assembly

With all the components meticulously prepared, we were ready to proceed to the crucial phase of system integration and assembly. This phase involves bringing together each individual component, ensuring they are correctly configured and interconnected to form a cohesive and fully operational system. The assembly process is essential not only for construction but also for verifying that all components function harmoniously as intended. This section details the step-by-step process of assembling and integrating all the components to create the final system, ready for comprehensive testing and validation.

This brief assembly process ensured that everything was properly connected and functioning as expected. To evaluate the performance of the final system before conducting the main survey, we generated a set of jokes to serve as test cases. These preliminary tests allowed me to observe the system's behavior and make any necessary adjustments to optimize performance ahead of the final evaluation.

During this phase, Meta released its latest model, Llama 3.1 [72], which quickly gained attention for its outstanding performance across various benchmarks. Despite being well into our development process, we decided to evaluate Llama 3.1 to see how it compared to our current system. We tested the Llama 3.1 model with 405B parameters using two different prompts and also fine-tuned the smaller Llama 3.1 model with 8B parameters using our custom dataset. The choice to fine-tune the 8B model instead of the larger 405B model was due to our computational limitations, as handling and fine-tuning a model of that size was beyond our available resources.

To compare our current system against these new models, we conducted a preliminary evaluation with five participants. Each participant was presented with four different jokes—one from each model—and was asked to choose their favorite. They also evaluated their favorite joke on a scale of 1 to 5 for funniness and explainability. This evaluation aimed to determine whether our system could generate jokes that resonated well with participants and whether we should consider switching to the Llama models for our final implementation.

The results were promising. Participants evaluated jokes based on 11 different news articles, resulting in 55 total evaluations. Of these, jokes generated by the Llama models were chosen only 7 times, indicating that our system outperformed the Llama models in the majority of cases. The jokes generated by our system achieved an average humor rating of 3.69, a significant improvement from the highest rating of 2.83 observed in our earlier comparison study.

While these initial results are encouraging and suggest that our system is effective in crafting humorous content, the small sample size limits the conclusiveness of the findings. Therefore, to draw more reliable conclusions, we plan to conduct a more extensive survey with a larger participant sample. This larger-scale questionnaire evaluation will provide the final insights needed to determine the overall effectiveness of our system and its potential for humor generation and explain news articles. Examples of the jokes generated by our system can be found in Appendix A for reference.

6

Questionnaire Evaluation

Contents

6.1	Structure	65
6.2	Results	68
6.3	Evaluation Discussion	89

Building upon the preliminary tests and system integration, the next pivotal phase involves a comprehensive evaluation of our humor generation system through a structured questionnaire. Humor, by its very nature, is highly subjective and varies widely across different audiences. Therefore, to accurately assess the effectiveness and appeal of the jokes generated by our system, it is imperative to gather insights from a diverse and substantial participant pool.

While informative, the preliminary evaluation was limited by a small sample size, which restricted the generalizability of its findings. To overcome this limitation and obtain a more nuanced understanding of our system's performance, we have designed a detailed questionnaire to capture participants' perceptions of humor, joke quality, and explainability.

6.1 Structure

In this section, we outline the structure of our questionnaire and explain the reasoning behind the questions we posed. The questionnaire was designed to explore the effectiveness of using humor as a tool to fight misinformation, a pressing issue in today's digital age. We begin by introducing the problem of misinformation and our innovative approach of using humor to address it. Following this, we describe the three types of exercises included in the questionnaire and the manner in which each exercise is presented to participants.

We began the questionnaire with a series of demographic questions to ensure that we could contextualize the survey results and identify any patterns based on participant backgrounds. Understanding the demographics of our participants allows us to analyze how factors such as age, gender, education level, and language proficiency might influence their perception of humor and its effectiveness in conveying information. The specific questions we asked were:

- What is your age group?
- What is your gender?
- What is your education level?
- What is your comfort level with the English Language, on a scale of 1-5?

These questions are important because age, education, and gender can all influence how individuals respond to humor and misinformation. Understanding these factors allows us to tailor our content more effectively for different audiences. Additionally, language proficiency is critical, as comfort with English can impact how well participants grasp jokes, particularly those that involve wordplay.

The next section of the questionnaire was designed to delve into participants' attitudes towards humor and their personal experiences with it. Understanding their general preference for and engagement

with humorous content is essential to gauge how likely they are to find humor effective in addressing misinformation. The questions in this section included:

- Do you enjoy jokes?
- How would you describe your typical reaction to humorous content?
- What types of jokes do you enjoy the most?
- How often do you search for jokes?
- How often do you share jokes with others?
- How often do you tell jokes to others?
- How do you typically find jokes?
- In your opinion, can jokes be used to effectively convey information or explain concepts?

This combination of demographic information and humor preferences sets the foundation for a comprehensive analysis of how different audiences perceive and engage with humor, particularly in the context of misinformation. By understanding these factors, we aim to draw more nuanced conclusions about the effectiveness of our approach and how it can be optimized for broader impact.

The next section of the questionnaire introduces the first type of exercise, where participants are presented with a joke accompanied by a claim, a label, and an explanation. It is important to mention that our survey has, not only machine generated jokes but also human-written ones, with the purpose of comparing the performance of both. For this type exercise, we ask a series of questions designed to gather comprehensive feedback on the joke's effectiveness and its impact on the participants. The questions are as follows:

1. Do you consider this explanation to be a Joke?
2. Do you think it is funny?
3. Do you consider it explains why the Claim was classified with that Label?
4. Do you consider it explains well?
5. To what extent do you think the joke could be offensive?
6. How engaging do you find the joke compared to the explanation itself?
7. Do you think you are more likely to remember the reason as to why the claim was classified with that Label?

8. How likely would you be to read the full article after seeing the joke, compared to if you only had seen the explanation?
9. How likely are you to share the joke compared to sharing the explanation?
10. How important is it for you to know the source of the information presented?
11. After reading the joke, to what extent do you feel it tried to manipulate you or the information in any way?

These questions are carefully designed to gather a wide range of insights. By exploring participants' reactions to the joke—such as whether they find it funny, memorable, or potentially offensive—we can draw meaningful conclusions about how humor influences the perception and retention of information. Additionally, questions regarding the likelihood of sharing the joke or seeking more information help us understand the broader impact of using humor in this context.

The second exercise in the questionnaire is similar to the first, but this time, participants are presented only with a joke and a claim. The jokes in this exercise include some from the previous one (questions one through six) and ask additional questions:

1. How would you classify this explanation?
2. If you believe the joke was created by a machine, what characteristics made you think so?
3. Which statement is better represented by the joke?
4. Do you think it is useful to include the source of the fact at the end of the joke?

This exercise aims to assess the joke's explainability from a different perspective. Instead of directly asking if participants think the joke explains well, we ask them to select the statement that best represents the joke's content. This approach offers an alternative view of the joke's explanatory power. Additionally, we inquire about the usefulness of including the source at the end of the joke, as this version includes a link to the original PolitiFact article from which the information was sourced at the end of the joke.

The final type of exercise in our questionnaire is a comparison exercise, which can take one of two forms: comparing two or three different jokes or comparing a joke, a poem, and a meme. In this exercise, we ask the following questions:

1. Which of these do you consider to be jokes?
2. Which one do you think is funnier?
3. Which one would you accept as an explanation?

4. Which format are you most likely to share with others?
5. If you encountered this claim online, which method would you prefer to read or engage with for clarification?
6. How does encountering Humor or Creative content (jokes, memes, poems) affect your emotional response to the information?
7. Do you think that creative content (jokes, memes, poems, etc.) enhances or detracts from the seriousness of the misinformation issue?
8. Do you think there are topics too serious to joke about?
9. Do you think creative content (jokes, memes, poems, etc.) simplifies complex information in a way that is helpful or misleading?

This exercise is designed to compare the effectiveness of jokes against other types of creative content and evaluate different humor approaches. Additionally, it allows us to compare jokes generated by a machine with those from our custom dataset. By exploring these key aspects, we can gather insights that further refine our understanding of how different types of content influence participants' engagement and perception. The survey used to conduct this evaluation is detailed in Appendix B, which also includes all the jokes presented to participants.

6.2 Results

To assess a wider range of jokes without making the survey overly long, we created three versions of the questionnaire. While each version contained the same set of questions, the news article presented varied between them. This approach allowed us to gather diverse data efficiently. We distributed the survey through various platforms, including Instagram, Discord, and WhatsApp, reaching a broad audience. In total, we collected 77 responses, with 21 participants responding to the first version, 30 to the second, and 26 to the third, ensuring that no participant responded to more than one version of the questionnaire.

Next, we will analyze the survey results in detail. Each question will be examined individually, followed by an analysis based on key factors such as age, educational background, and language proficiency.

6.2.1 Demographic and Humor preferences Data

We begin by presenting the demographic data of the questionnaire participants. In terms of gender distribution, 50.7% of the 80 respondents were female, while 49.3% were male. Regarding age groups, the

largest portion (34.3%) consisted of participants aged 18 to 25. The second-largest group, representing 24.3%, fell within the 51 to 65 age range, followed by participants aged 26 to 35, at 21.4%. The smallest group, comprising 20% of participants, was between 36 and 50 years old. In terms of educational background, 12.7% of respondents held a high school diploma, 43.7% had earned a Bachelor's degree, 35.2% had a Master's degree, and 8.5% held a PhD. Language proficiency was rated on a scale from 1 to 5, where 1 indicated "uncomfortable but able to read in English" and 5 represented "native speaker". Almost half of the participants (49.3%) rated their proficiency as 4, while 28.2% selected 3. A smaller group of 19.7% rated themselves as native speakers (5), and only 2.8% rated their proficiency at level 2. The detailed demographic data can be found in Figure 6.1.

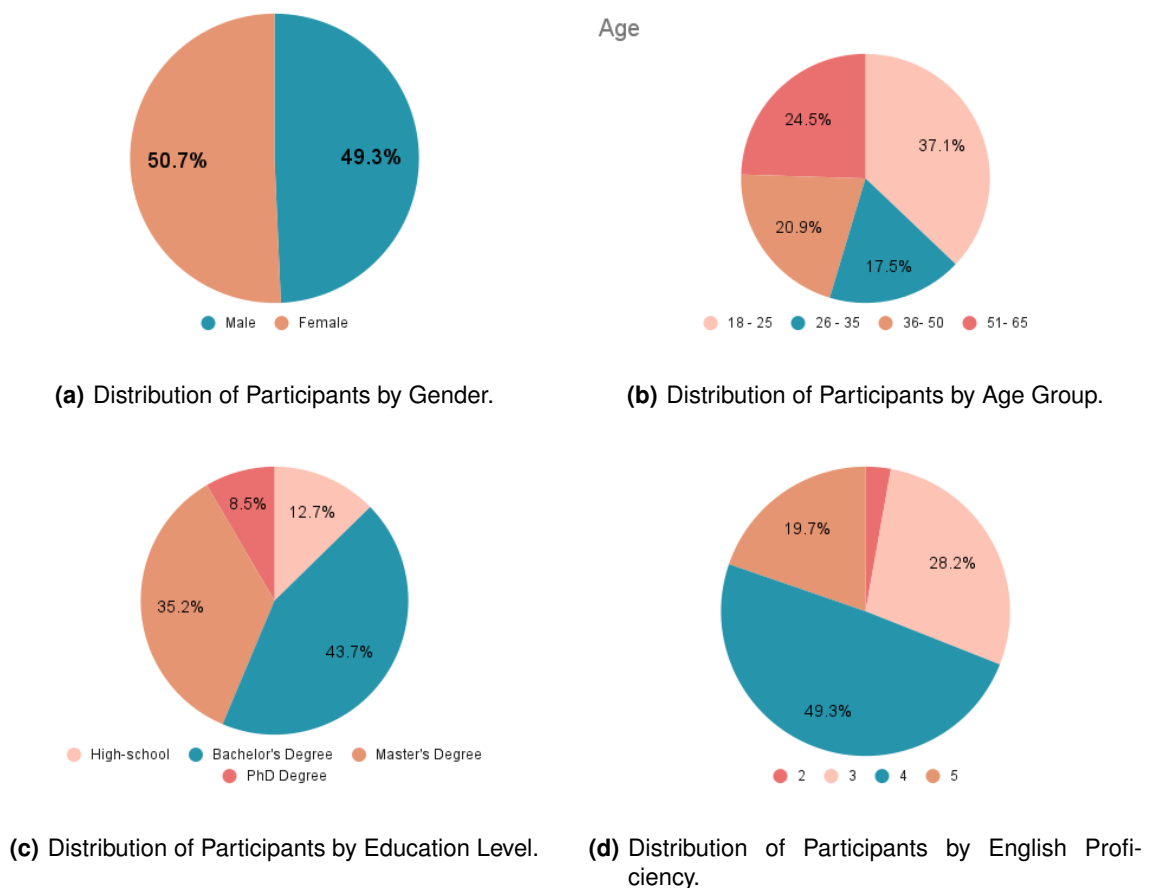


Figure 6.1: Demographic Data.

In addition to demographic data, we also gathered information on the humor preferences of participants. The first question, "Do you enjoy jokes?" received a unanimous 100% "Yes" response. We then asked participants about their typical reaction to humorous content, with the top responses being "Laughter" (80.3%) and "Amusement" (70.4%). When it came to favorite types of jokes, dark humor topped the list at 71.8%, alongside satire also at 71.8%. Puns, one-liners, and anecdotes were close

in percentage, receiving 50.7%, 46.5%, and 45.1%, respectively. Regarding how frequently participants seek out jokes, 19.7% reported searching for jokes multiple times a day, 36.6% do so once a day or less, and 25.4% look for jokes once a month or less. The rest of respondents (18.3%) selected “Never”. The next question asked how often participants share jokes with others. A notable 77.5% reported sharing jokes at least once a week, a trend mirrored by 73.3% who stated that they tell a joke more than once a week. Regarding where participants find jokes, the top three sources were social media (83.1%), TV series and movies (69%), and messaging apps such as WhatsApp (59.2%). The final question focused on whether participants believed humor could be used to explain concepts. A promising 62% answered “Yes”, while the remainder selected “Sometimes”, indicating a generally positive outlook for our research.

6.2.2 Single Exercises

We will examine the results of the human-written jokes first and then proceed to examine the machine-generated ones. We will go through every question mentioned for the first two types of exercise.

6.2.2.A Human Written Jokes

The first question participants answered was, “Do you consider this explanation to be a joke?” The results, shown in Figure 6.2, indicate that 70.2% of respondents chose “Yes” or “Definitely Yes,” while 11.1% selected “No” or “Definitely No”. The remaining 18.7% opted for “Maybe”.

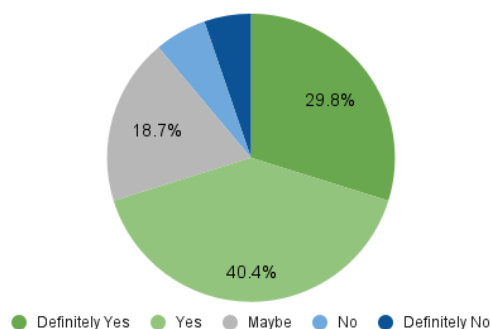
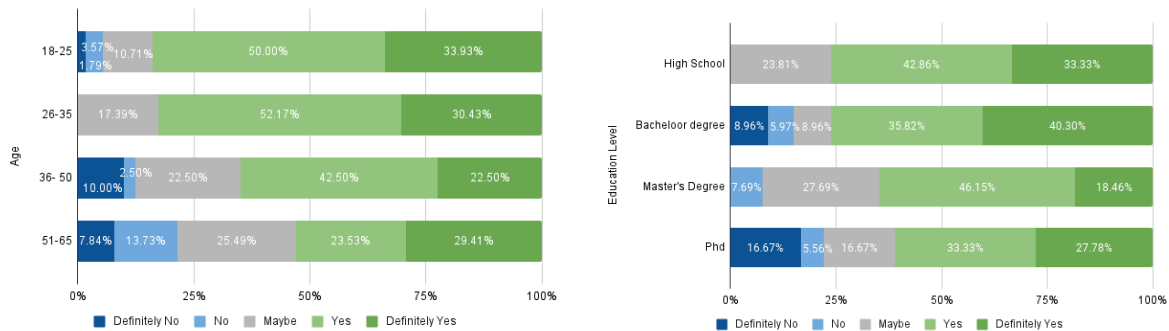


Figure 6.2: Results distribution of “Do you consider this explanation to be a joke?” for human-written jokes.

When analyzing these responses by age group (Figure 6.3(a)), we observe that participants aged 18 to 25 had the highest percentage of people who considered the explanation to be a joke, with 83.9% selecting “Yes” or “Definitely Yes”. As age increases, participants tend to be more critical in determining whether the explanation qualifies as a joke. While the “Yes” and “Definitely Yes” percentages remain relatively consistent across the older age groups, the proportion of “No” and “Definitely No” responses steadily increases, peaking in the 51 to 65 age group. A similar trend is evident when analyzing

responses based on education level (Figure 6.3(b)). The higher the level of education, the fewer participants selected “Yes” or “Definitely Yes”. Conversely, the proportion of “No” and “Definitely No” responses increases, starting at 0% among participants with a high school diploma and reaching its highest point at 22.2% among those with a PhD.



(a) Results distribution based on Age.

(b) Results distribution based on Education Level.

Figure 6.3: Results distribution of “Do you consider this explanation to be a joke?” for human-written jokes.

Next, we analyzed the ratings given to the jokes. Human-written jokes received an average rating of 3.5 on a 1-5 scale. To further explore the ratings, we visualized their overall distribution using a box chart (Figure 6.4). The chart reveals that the median rating is 3.5, aligning with the average rating. The interquartile range spans from approximately 3.0 to 4.0, indicating that the majority of responses are clustered within this range. The whiskers extend from about 2.0 to 5.0, reflecting the broader variability in ratings. Notably, a single outlier below 2.0 is present, indicating that while most participants rated the jokes favorably, a small minority found them significantly less appealing. Overall, the box chart underscores that the ratings are moderately concentrated, with a general tendency toward positive evaluations.

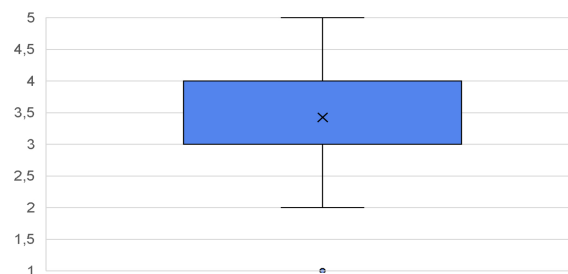


Figure 6.4: Box chart illustrating the distribution of funniness ratings for human-crafted jokes.

When broken down by age, the ratings tend to decline with increasing age. Participants aged 18 to 25 gave the highest average rating of 3.7, which drops to 3.4, then 3.3, and finally 3.1 for the older age groups. This suggests that since people aged 18 to 25 created the jokes, they may also resonate

more with people in that age range, possibly due to shared cultural references, humor styles, or generational perspectives. Regarding education level, participants with a high school diploma gave the highest average rating of 3.9. This rating decreases slightly to 3.5 among those with a Bachelor's degree, and continues to decline to 3.3 and 3.2 for participants with a Master's and PhD, respectively. This pattern mirrors the trend observed in the age analysis, reinforcing the conclusion that jokes tend to resonate more strongly with people within the same age range as those who created them. To further illustrate these trends, the box charts, present in Figure 6.5 were created for both age and education level data. These visualizations provide a clearer picture of the variability and distribution of ratings within each group. These visualizations provide a clearer picture of the variability and distribution of ratings within each group, highlighting the general decline in ratings across increasing age and education levels.

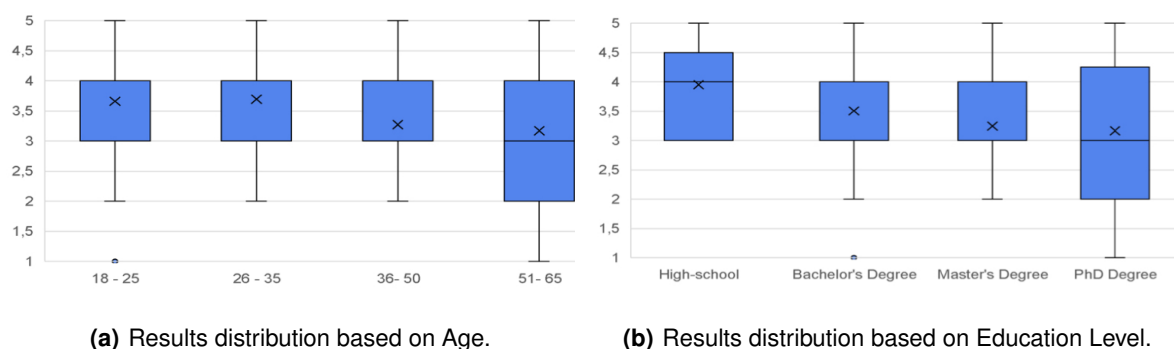


Figure 6.5: Box chart illustrating the distribution of funniness ratings for human-crafted jokes.

We then analyzed the explainability of the jokes, focusing on whether participants believed the joke effectively explained why a particular claim was classified with a specific label. The results were generally positive, with 67.8% of respondents selecting “Yes” or “Definitely Yes”, and only 7.6% responding with “No” or “Definitely No”. The remaining 24.6% chose “Maybe”. This positive sentiment was consistent across all age ranges, with over 50% of respondents in each group selecting “Yes” or “Definitely Yes”. However, there was a slight increase in “No” and “Definitely No” responses as age increased. For participants aged 18 to 25, only 3.6% selected negative responses. This percentage rose to 4.4% in the 26 to 35 age group, 6.8% in the 36 to 50 range, and peaked at 17.7% in the 51 to 65 group. This trend suggests that older participants were more critical of the jokes’ ability to explain the claim. A similar pattern emerged when analyzing the data by education level, with one notable exception. While most education levels mirrored the age trends—showing a majority of positive responses and a gradual increase in critical responses as education level rose—participants with a PhD deviated from this pattern. Only 38.9% of PhD holders responded with “Yes” or “Definitely Yes”, while the majority, 44.4%, selected “Maybe”. This suggests that participants with a PhD were more uncertain about the explanatory power of the jokes, potentially indicating that they found the humor lacking in depth for more complex explanations.

However, despite this uncertainty, when asked to rate how well the joke explained the claim on a

scale of 1 to 5, the average rating remained consistent at 3.6 across all age and education levels, with a slight variation of only 0.2, as we can see in Figure 6.6. This consistency suggests that while PhD participants may have been more critical in choosing clear affirmative or negative responses, they still rated the overall clarity similarly to other groups. In other words, the hesitation reflected in their “Maybe” responses might indicate a desire for more strict explanations, even though they did not find the clarity of the humor significantly worse compared to other educational backgrounds.

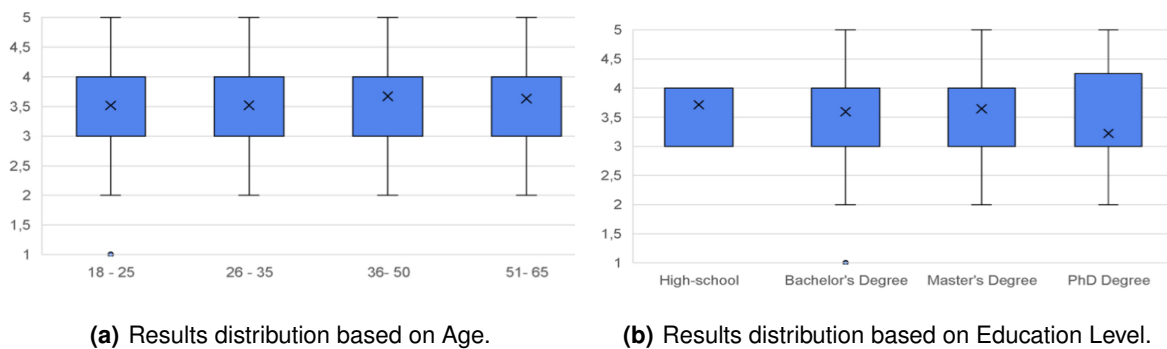


Figure 6.6: Box chart illustrating the distribution of explainability ratings for human-written jokes.

Regarding the offensiveness of the human-written jokes, participants rated the jokes on a scale of 1 to 5, where 1 indicates “not offensive” and 5 indicates “very offensive”. The overall average rating for offensiveness was relatively low, at 1.5. However, we observed a noticeable trend when breaking the data down by age and education level. For age, there was a gradual increase in the perceived offensiveness of the jokes. Participants between 18 and 35 rated the jokes at an average of 1.2, while those between 36 and 65 rated them higher, with an average of 1.8. This suggests that younger participants tend to perceive the jokes as less offensive, possibly due to generational differences in humor tolerance or shifting social norms. Older participants, on the other hand, might have found certain elements of the humor more inappropriate or less aligned with their sensibilities, leading to a higher offensiveness rating. A similar pattern emerged across education levels. Participants with a high school diploma or a Bachelor’s degree rated the jokes at 1.2 on average, while those with a Master’s or PhD rated them higher, at 1.8. This could reflect differences in the expectations for humor across educational backgrounds. Higher-educated participants might be more sensitive to nuances in humor that could be perceived as offensive, or they may have different standards for what they consider appropriate in a joke.

The final question we analyzed focused on participants’ perceptions of the joke’s origin. They were asked whether they believed the joke was human-written or machine-generated, followed by a request to explain their reasoning if they thought it was machine-generated. A notable 31.7% of participants were uncertain about the joke’s origin, indicating that distinguishing between human and machine-generated jokes was not a straightforward task for many. Among those who did make a choice, 42.3% believed the

joke was human-written, while 26.1% thought it was machine-generated.

While a higher percentage correctly identified the human origin, the significant proportion of participants who were either unsure or misidentified the joke as machine-generated offers valuable insight. The fact that nearly a third of respondents were uncertain suggests that, to many, the boundary between human and machine-generated humor is increasingly indistinct. This blurring of lines could be seen as a positive outcome for our system, as it indicates that machine-generated jokes are convincingly human-like to some participants. However, it also underscores the need for further refinement.

For the 26.1% of participants who believed the jokes were machine-generated, several key reasons emerged. Common feedback included a perceived lack of humor, unusual structure or grammar, and a sense that the joke was “emotionless” or lacked the nuanced delivery typically associated with human creativity. These responses highlight areas for improvement. A lack of humor or awkward phrasing points to potential weaknesses in the jokes we made to train our system. These insights also suggest an opportunity to refine the dataset of human-written jokes. If these human-created jokes can be improved in terms of humor, structure, and relatability, it may further enhance the performance of the machine-learning model by providing better examples from which to learn.

Next, we will analyze the performance of machine-generated jokes, focusing on their perceived humor quality and their relevance to our study. The remaining, more general questions will be addressed afterward, as they are intended to assess the overall effectiveness of humor in fighting misinformation, rather than comparing human-generated and machine-generated jokes.

6.2.2.B Machine Generated Jokes

We analyze the questions in the same order as in section 6.2.2.A. Regarding machine-generated jokes, participants rated them as jokes 6.7% more frequently than human-written jokes. Specifically, 76.9% of respondents answered “Yes” or “Definitely Yes”, as illustrated in Figure 6.7. When examining the data by age, we observe a trend similar to that of human-made jokes: the percentage of negative responses increases with age. Specifically, the percentage of “No” and “Definitely No” responses rises from 2.3% in the 18 to 25 age group to 8.8% in the 51 to 65 age group. In contrast, the proportion of positive responses remains relatively stable across all age ranges, as shown in Figure 6.8(a). In the analysis of education level, Figure 6.8(b) indicates that the percentage of positive and negative responses is nearly uniform among participants with education levels higher than high school. Notably, the high school group stands out with 0% negative responses, suggesting that this demographic may have a more favorable view of humor.

Interestingly, despite human-made jokes serving as the training data for a part of our model, the machine-generated jokes outperformed them in terms of positive responses. This improvement may be attributed to the three distinct methods employed in generating jokes, allowing for a greater variety

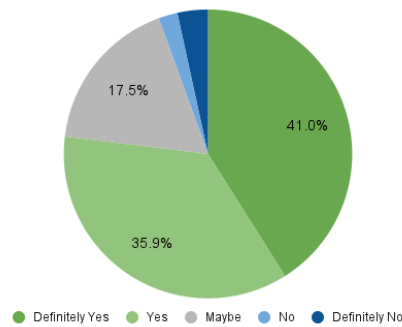
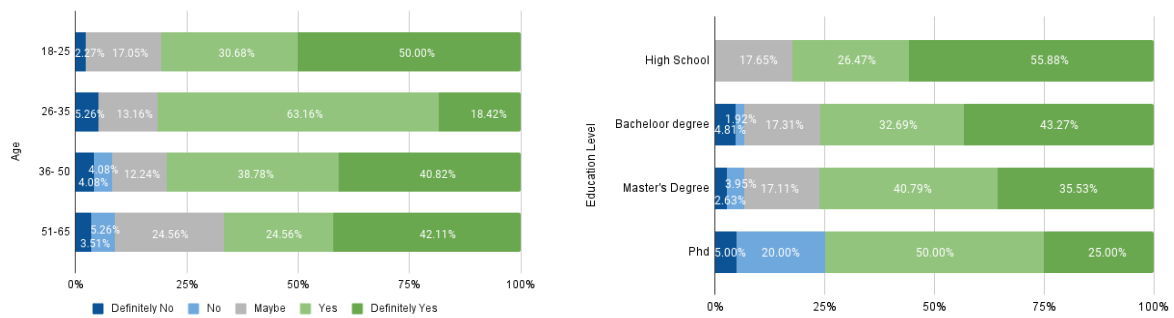


Figure 6.7: Results distribution of “Do you consider this explanation to be a joke?” for machine-generated jokes.



(a) Results distribution based on Age.

(b) Results distribution based on Education Level.

Figure 6.8: Results distribution of “Do you consider this explanation to be a joke?” for machine-generated jokes.

and creativity in humor that resonates more effectively with participants. This data implies that while machine-generated jokes are generally well-received, perceptions of humor do vary by age and education level. The steady increase in negative responses among older participants may suggest a more critical attitude toward humor as individuals age. Conversely, the lack of negative responses from those with only a high school education may reflect a greater tolerance or appreciation for humor, regardless of its source.

The jokes achieved an average funniness rating of 4, indicating that participants generally found them amusing, as illustrated in Figure 6.9. Interestingly, this funniness rating remained remarkably consistent across different age ranges, with only slight fluctuations of approximately 0.1. As shown in the boxplot in Figure 6.10(a), participants aged 18–25 provided slightly higher ratings, while those in the 26–35 and 51–65 age groups rated the jokes slightly lower. Nonetheless, the average rating across all age groups hovered around 4, suggesting a broad consensus on the jokes’ humor regardless of age. A more pronounced variation emerges when analyzing funniness ratings by education level. As depicted in Figure 6.10(b), participants with a high school education rated the jokes the highest, with a mean score of 4.5. In contrast, individuals with higher levels of education provided progressively lower ratings,

with PhD holders assigning an average score of 3.8. The increasing number of ratings below 4 among those with more advanced education suggests they may have been more critical or had different humor preferences.

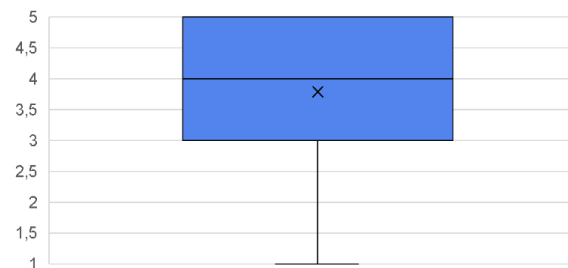


Figure 6.9: Box chart illustrating the distribution of funniness ratings for machine generated jokes.

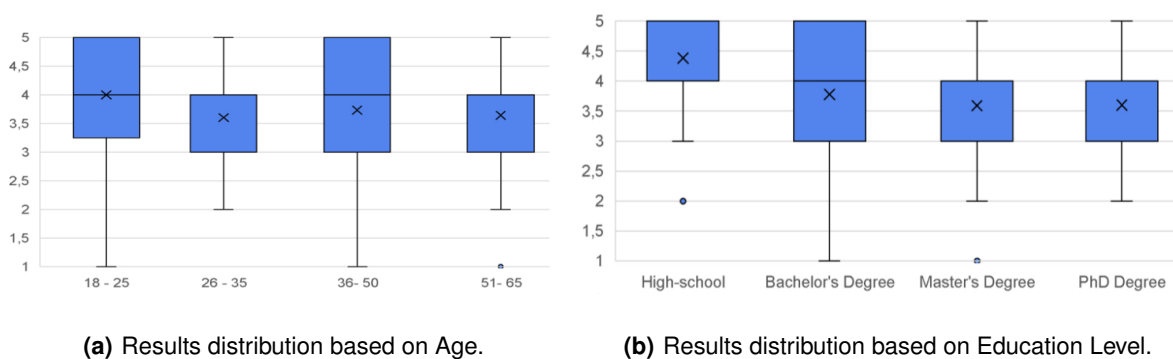


Figure 6.10: Box chart illustrating the distribution of funniness ratings for machine generated jokes.

Regarding the explainability of the jokes, we saw positive feedback, with 74.7% of participants providing favorable responses, while only 9.9% gave negative feedback, and 15.4% were uncertain. However, when analyzing these responses by age, we could not draw strong conclusions, as the percentage of positive responses fluctuated. For example, the percentage dropped from 74.7% in the 18-25 age group to 68.4% in the 26-35 group, before increasing in older participants. On the other hand, the data showed a clearer trend when broken down by education level. Participants with a high school degree provided the highest percentage of “Definitely Yes” responses, implying that individuals with lower education levels found the explanations more satisfactory. As education level increased, participants seemed to demand more rigorous explanations, although the overall percentage of positive responses remained high. This suggests that while the jokes were explainable to a broad audience, those with more education sought more depth in the explanations.

One of the key insights from our study came from the ratings related to explainability, which averaged 3.9 overall. These ratings remained fairly stable across age groups, but there was more variation based on education levels. Participants with high school or PhD degrees gave the highest ratings, at 4.4 and 4.1,

respectively, while participants with other degrees rated explainability lower, at 3.7 and 3.8. This result was surprising, as PhD holders had previously been among the most critical, yet they rated explainability relatively high. This suggests that while these participants were strict in evaluating funniness, they valued the clarity of the explanations. The box charts in Figure ?? illustrate the ranges of ratings given by participants, further highlighting these trends and offering a visual confirmation of the patterns observed in the data.

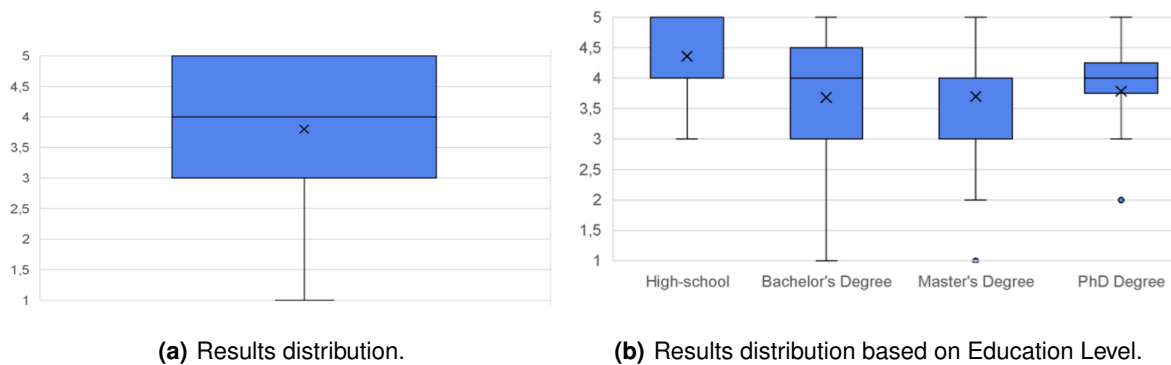


Figure 6.11: Box chart illustrating the distribution of explainability ratings for machine generated jokes.

Another important question examined whether participants were able to guess the underlying message of the jokes. Out of the 76 responses we received, 57 participants (75%) correctly identified the intended statement, while 9 participants (11.8%) guessed incorrectly, and 10 participants (13.2%) did not select any of the options. Although most participants managed to identify the correct message, the 26.8% who either guessed incorrectly or did not choose an option suggests a potential trade-off between humor and explainability. It is possible that, in an effort to make the jokes funnier, some clarity in conveying the intended message was sacrificed, despite overall success in humor and comprehension.

On the topic of offensiveness, the machine-generated jokes received a low average rating of 1.5, indicating that participants found them largely inoffensive. While this score did not represent a decrease compared to human-written jokes, it suggests that our moral classifier is effectively filtering out potentially offensive content. Unlike human-written jokes, where offensiveness varied significantly across age groups, the ratings for machine-generated jokes were fairly uniform, ranging only from 1.4 to 1.5. This consistency implies that the moral standards of the jokes were well-balanced across age groups, thanks to the classifier. However, there was a notable variation based on education level. Participants with a high school degree gave the lowest offensiveness rating of 1.2, while participants with bachelor's, master's, and PhD degrees gave higher ratings, ranging from 1.6 to 1.7. This could indicate that individuals with more education may have been more sensitive to certain elements of the jokes, even if they were not overtly offensive.

Lastly, we considered participants' perceptions of the joke's origin—whether they thought the joke

was human-written or machine-generated. A significant percentage of participants (25.6%) were unsure of the origin, while 44.7% believed the joke was human-written. This represented a small increase of 2.7% compared to human-written jokes, which is a promising sign that our system can generate jokes that are perceived as more human-like. However, there was also a 3.8% increase in the number of participants who thought the joke was machine-generated, citing reasons such as lack of humor, being overly explanatory, or perceiving parts of the explanation as copied or repetitive. While these criticisms indicate areas for improvement, the fact that nearly half of the participants believed the jokes were human-made shows that our system is successfully challenging preconceived notions about machine-generated humor.

Overall, the results suggest that machine-generated jokes are well-received, often surpassing human-written jokes in terms of humor, explainability, and even perceived origin. This indicates strong performance from our system. However, there is potential for further improvement, particularly by refining our dataset of human-written jokes. By enhancing the funniness and variety of the dataset, we may be able to further boost the system's ability to generate jokes that appeal to an even broader audience, and possibly achieve higher levels of both humor and explainability.

6.2.2.C Effectiveness of Jokes in Fighting Misinformation

In this section, we analyze whether humor, specifically in the form of jokes, can serve as an effective tool to counteract the rapid spread of misinformation. We first examine responses to the question: "How engaging do you find the joke compared to the explanation itself?" Participants rated their engagement on a scale of 1-5, where 1 indicated "Much less engaging" and 5 indicated "Much more engaging". The distribution of responses reveals interesting insights, as shown in Figure 6.12. A significant portion of participants (29.1%) rated jokes as "much more engaging" than textual explanations, while 36.3% found jokes to be "more engaging". Combined, this suggests that 65.4% of respondents considered humor to be a more engaging medium than straightforward factual explanations. In contrast, 24.6% viewed the joke and the explanation as equally engaging, while only 9.9% found the joke less engaging than the textual explanation. Upon closer analysis, we observe that lower engagement scores for jokes were correlated with lower ratings of the joke's humor. This indicates that the effectiveness of humor as a tool to engage depends significantly on the quality or funniness of the joke. When the joke resonates and is deemed funny, it tends to outperform or at least match the engagement levels of factual text.

These findings highlight the potential of humor to serve as a useful tool in fighting misinformation. Since misinformation often spreads quickly by leveraging emotional appeal and engagement, integrating humor into factual content could increase its appeal and make accurate information more shareable. Therefore, humor, when executed effectively, can enhance the presentation of factual information and serve as a counter-strategy to the emotional allure of misinformation.

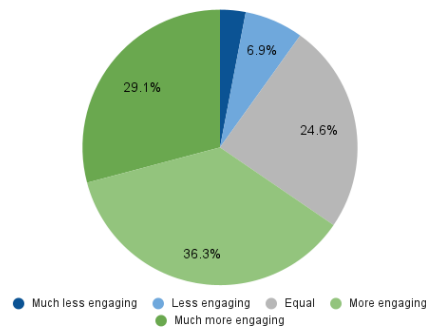
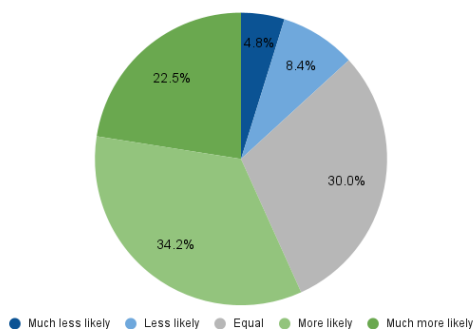
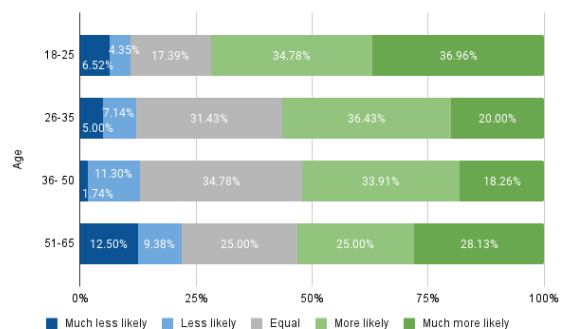


Figure 6.12: Results distribution of “How engaging do you find the joke compared to the explanation?”.

The next question focused on whether jokes could improve participants’ ability to retain the reasons behind why a specific claim was labeled in a particular way (e.g., false, half-true, etc.). Our goal was to assess whether humor could enhance information retention, potentially making people more likely to remember the rationale behind the classification. The results, shown in Figure 6.13(a), reveal that 56.7% of participants believed that jokes improved their likelihood of remembering why the claim received its label. Meanwhile, 30% felt their retention remained unchanged, and 13.2% believed their retention worsened when jokes were used. When breaking down these results by age groups (as shown in Figure 6.13(b)), an interesting trend emerges: as age increases, the percentage of participants who think that jokes enhance information retention decreases. Conversely, the proportion of participants who feel that jokes negatively impact retention increases with age. Despite the rise in negative responses among older participants, the overall findings remain promising. More than half of the participants (56.7%) reported that jokes helped them better remember the reasons for a claim’s label. This is a significant result, as the goal is not only to convey factual information but also to ensure that audiences remember why a piece of misinformation is false or misleading.



(a) Results distribution.



(b) Results distribution based on Age.

Figure 6.13: Results distribution of “Do you think you are more likely to remember the reason as to why the claim was classified with that Label?”.

Next, we explored whether reading a joke about a claim made participants more likely to read the full article containing a detailed explanation of why the claim received a particular label. This question is critical because while jokes can engage people, they often cannot provide all the necessary details. Therefore, it is important to encourage readers to seek out the full factual explanation. The results, displayed in Figure 6.14, show that 40.5% of participants reported being more likely to read the full article after reading the joke. However, 29.4% felt the joke had no impact on their decision, and 30% said they would not read the full article. Interestingly, age played a notable role in shaping these responses. Among participants aged 18 to 25, a higher percentage (39.2%) indicated they would not read the full article after reading the joke, compared to 37.5% who said they would. This is the only age group where the percentage of those unlikely to seek further information exceeds those who are likely to do so. For older age groups, the trend shifts more positively, with participants being more inclined to pursue detailed information after engaging with a joke.

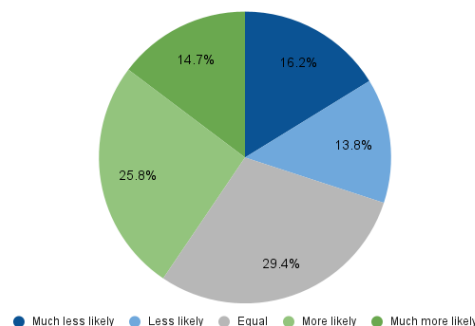


Figure 6.14: Results distribution of “How likely would you be to read the full article after seeing the joke?”.

These findings suggest that while humor can serve as an effective hook to capture attention, particularly for older audiences, its impact on driving deeper engagement with factual content varies by age. The younger demographic, particularly those aged 18 to 25, appear more hesitant to seek out additional information after reading the joke, which may point to a preference for shorter, more digestible content. Despite this, the overall result is still promising, with a plurality (40.5%) of participants across all age groups expressing an increased likelihood of reading the full article after encountering the joke. This reinforces the idea that humor can serve as a bridge to more thorough explanations, encouraging readers to explore factual information beyond the surface level. However, to maximize its effectiveness, it may be necessary to adapt the format or presentation of the follow-up content, especially for younger audiences who may require different strategies to drive further engagement.

One key factor in fighting misinformation is reducing its spreadability. With this in mind, we asked participants how likely they would be to share a humorous explanation (joke) compared to a traditional textual explanation. The goal was to determine whether humorous content could achieve higher shareability, thereby amplifying the reach of factual information. According to the results, present in Figure 6.15,

57.7% of participants stated they would be more likely to share the joke than the textual explanation. Additionally, 20.7% indicated they would be equally likely to share either, while the remainder were less inclined to share the joke. When analyzing these responses by age and education level (Figures 6.16(a) and 6.16(b)), a clear pattern emerges: as both age and education increase, the likelihood of sharing the joke declines. The lowest rate of sharing was found in the 51–65 age group, where only 42.9% reported being more likely to share the joke, and this group also had the highest percentage of participants (25.2%) who indicated they would be less likely to share it. This decline in shareability among older and more educated participants is consistent with previous findings in the survey, where older age groups generally reported sharing jokes less frequently overall. Therefore, while humor seems to be an effective tool for encouraging information sharing among younger and less formally educated audiences, it may be less effective with older and more educated demographics.

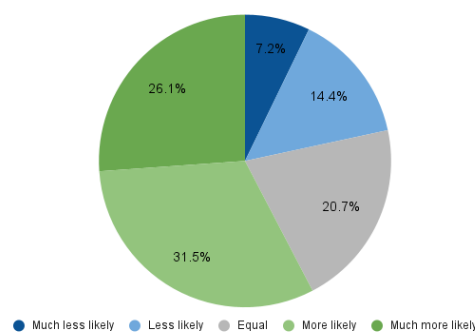


Figure 6.15: Results distribution of “How likely are you to share the joke?”.

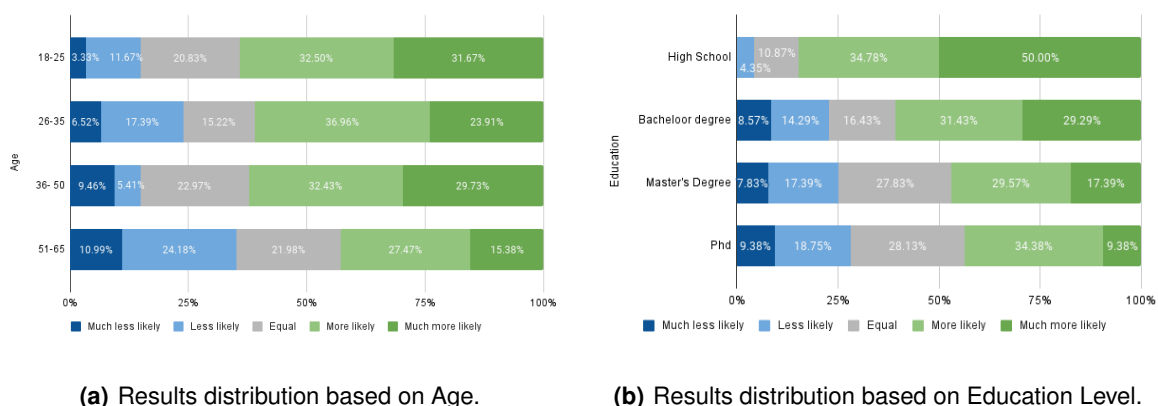


Figure 6.16: Results distribution of “How likely are you to share the joke?”.

The next question aimed to assess how important participants found it to know the source of the information presented to them. Responses were rated on a scale from 1 to 5, where 1 indicated “Not important at all” and 5 indicated “Extremely important”. The results of this question were concerning, as

31% of participants selected 1 or 2, indicating low importance for knowing the source. However, 51% of participants rated the source's importance as either a 4 or 5, showing that a slight majority still placed significant value on source credibility. When we break down the data by age, an interesting pattern emerges. The age group of 18 to 25 showed the highest proportion of negative responses, with 40.3% of participants in this group rating the importance of knowing the source as low (1 or 2), as shown in Figure 6.17. This is a concerning result, as it indicates that younger individuals may be more indifferent to the credibility of information sources, making them potentially more vulnerable to misinformation. In contrast, as age increases, the percentage of participants who consider the source to be important also increases. Older age groups tend to assign higher importance to source credibility, although there are still relatively high percentages of participants in these groups who do not consider it as crucial. This trend suggests that older individuals are more likely to value the source when evaluating information, but the persistent skepticism or apathy across all age ranges points to a broader issue regarding media literacy. The fact that a substantial number of participants, particularly among younger demographics, do not prioritize the source of the information they receive poses a significant challenge in the fight against misinformation. If people are indifferent to where their information comes from, they are more likely to accept and share false or misleading claims without critical evaluation. This undermines efforts to spread accurate information, as misinformation often spreads more quickly and widely due to its emotional appeal, regardless of the credibility of its source. To address this, improving media literacy and emphasizing the importance of source verification should be key components of any strategy aimed at fighting misinformation. While humor and engaging content can help capture attention, ensuring that individuals recognize and value credible sources of information is equally critical to reducing the spread of false narratives.

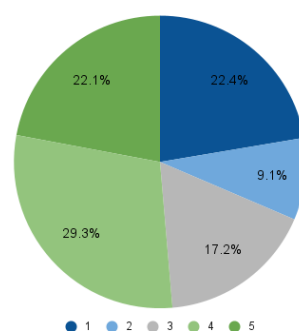
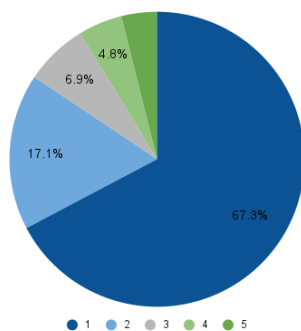


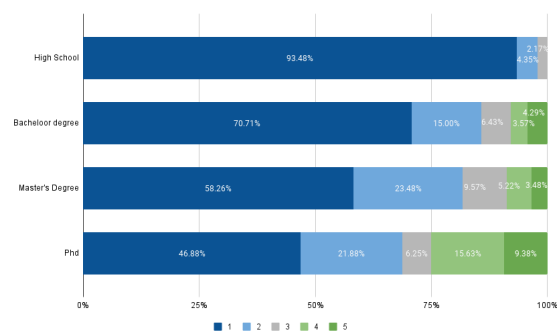
Figure 6.17: Results distribution of “How important is it for you to know the source of the information presented?”, (1 = “Not important at all”, 5 = “Extremely important”).

Lastly, we examined whether participants felt that our jokes were attempting to manipulate them or distort the information presented. This is a critical issue, as one of our main goals is to fight misinformation without resorting to manipulative tactics. The results, illustrated in Figure 6.18(a), show that only

a small percentage of participants believed the jokes were manipulating the information. However, as the education level of participants increased, so did the perception of manipulation, as shown in Figure 6.18(b). This perception peaked at 25% among those with the highest education levels. The low overall percentage of participants who felt manipulated is encouraging, as it suggests that the majority of respondents recognized the intent of the jokes as informative rather than deceptive. This is crucial, as humor should aim to engage and clarify rather than mislead. However, the increase in perceived manipulation among those with higher education levels warrants closer examination. It's possible that individuals with more formal education are more critical of content and may be more sensitive to nuances in tone or presentation, leading them to question the authenticity or objectivity of humorous material. To mitigate this, it is important to ensure transparency and maintain a balance between humor and factual integrity. Providing clear sources or additional context alongside the jokes may help alleviate concerns of manipulation, especially for more educated audiences who value critical analysis and depth. Overall, while the general response indicates that most participants did not view the jokes as manipulative, it is essential to remain mindful of how different audiences perceive humor in order to maximize its effectiveness without compromising trust.



(a) Results distribution.



(b) Results distribution based on Age.

Figure 6.18: Results distribution of “After reading the joke, to what extent do you feel it tried to manipulate you or the information in any way?” (1 = “Not at all”, 5 = “Extremely”).

6.2.3 Comparison Exercises

In this section, we analyze the results of three comparison exercises designed to evaluate the effectiveness of different humor-based approaches for fighting misinformation.

The first exercise compares three jokes: one generated by our system, one created during the experiment described in section 5.2.2.E, and a third produced by the default GPT-4 model. The goal is to determine whether our domain-specific fine-tuning leads to better jokes compared to our experimental

system and GPT-4's baseline model. This comparison helps assess the impact of fine-tuning on the quality of humor and its effectiveness in conveying factual information.

The second exercise involves a comparison between a joke generated by our system and one written by a human. The purpose here is to evaluate both the funniness and the explanatory strength of the jokes, helping us understand how well our system performs compared to human-generated humor in terms of engagement and clarity.

The final exercise compares different forms of creative content—jokes, memes, and poems—to determine which method is the most effective for explaining information and appealing to the audience. Each creative format is also compared to a traditional textual explanation to gauge its overall impact on understanding and engagement.

We begin by analyzing the results of the first exercise where the first question asked was “Which of these do you consider to be jokes?” where participants could select all the options they deemed as jokes. The results show that 87.3% of participants considered the joke generated by our system to be a joke. This is a strong indication that the humor produced by our system is largely recognized as effective, validating its capacity to create content that aligns with people's expectations of what constitutes a joke. Next, 70.4% of participants identified the joke produced by the domain-specific fine-tuned model as a joke. While this result is lower than the joke from our system, it still demonstrates promising potential. The fact that more than two-thirds of respondents recognized it as a joke suggests that fine-tuning can improve joke quality in domain-specific contexts, but may require further refinement to reach the same level of engagement as our final system. In contrast, only 38% of participants considered the joke generated by the default GPT-4 model to be a joke. Although this is significantly lower than the other two, it is still noteworthy because this output was created by the base, unmodified GPT-4 model. The relatively low recognition rate suggests that while GPT-4 can generate humorous content without domain-specific tuning, it struggles to produce jokes that align with audience expectations in a highly targeted context. This result highlights the importance of fine-tuning and different approaches in specific tasks, such as humor generation. Lastly, a small 4% of participants indicated that none of the explanations presented were jokes. This minority shows that humor perception varies among individuals, and some participants may have more rigid expectations of what constitutes a joke. It also underscores the subjective nature of humor, where certain formats or styles may not resonate with everyone. Overall, these results underscore the effectiveness of our system in generating jokes that are widely recognized as such, while also highlighting the importance of fine-tuning and contextual adjustments to maximize the humor creation.

The follow-up question asked participants which of the three jokes they found to be the funniest. The results show that 69% of participants selected the joke generated by our system as the funniest, while 14.1% chose the joke generated by the domain-specific fine-tuned model, and 11.3% found the

joke generated by the default GPT-4 model to be the funniest. The remaining participants opted for “None” as their answer. These results are quite revealing. First, the overwhelming preference for the joke generated by our system (69%) suggests that it not only meets participants’ expectations for what constitutes a joke, but also resonates well in terms of humor. This reinforces the notion that our system is effective at producing jokes that a majority of people find amusing and engaging. Interestingly, only 14.1% of participants found the fine-tuned model’s joke to be the funniest, despite 70.4% of respondents previously recognizing it as a joke. This gap between recognition and humor indicates that while the domain-specific fine-tuning successfully produces content that participants identify as jokes, it may lack the comedic appeal needed to make them laugh. This suggests that although the fine-tuned model performs reasonably well in creating domain-relevant humor, it struggles to match the quality of humor generated by our system. Similarly, the default GPT-4 model was selected as the funniest by 11.3% of participants, slightly lower than the fine-tuned model. While this percentage is modest, it’s notable that the base GPT-4 model—without any domain-specific tuning—still managed to elicit humor from some participants. However, the relatively low figure emphasizes that GPT-4, in its default form, may not be as effective at generating widely appealing humor, particularly in comparison to a more tailored system. Finally, a small portion of participants chose “None” as their answer, indicating that they did not find any of the jokes particularly funny. When examining the data across different age groups and education levels, the trend remains consistent—our system’s joke was favored as the funniest across all demographics, though the exact percentages varied. This uniformity across diverse participant groups suggests that our system has broad appeal, successfully generating humor that resonates with a wide range of individuals.

The final question in this exercise asked participants which of the humorous explanations they considered to be an actual explanation. The results were less favorable in this case. Only 22.5% of participants regarded the joke generated by our system as an explanation, while 18.3% considered the fine-tuned model’s joke to serve as an explanation. In contrast, 59.2% of participants identified the output generated by the default GPT-4 model as an explanation.

These results highlight a critical trade-off between humor and explanatory clarity. This suggests that, despite the positive results in the section 6.2.2.B, as we increase the humor in our content, we may be compromising the depth and clarity required for participants to perceive it as an actual explanation when compared to a more formal explanation. The relatively low percentages for both our system and the fine-tuned model reflect this challenge, where the humor may overshadow the educational value, reducing the overall clarity of the message. Interestingly, the default GPT-4 model, despite being less engaging in terms of humor, was considered a more effective explanation by a majority of participants (59.2%). This suggests that while GPT-4 may lack humor, it retains a more straightforward explanatory structure that participants are more likely to accept as informative. The GPT-4 output likely adheres more closely

to traditional explanatory formats, which may make it easier for participants to understand the factual information without the distraction or ambiguity introduced by humor. The results highlight an important balance between engagement and clarity. While humor can capture attention and make content more appealing, there is a risk of diluting the informational content to the point where participants no longer view it as a legitimate explanation. This trade-off is important when considering the overall goals of fighting misinformation: while humor can enhance engagement, it must be carefully balanced to ensure that the factual information is still clearly communicated and recognized as such.

We now turn our attention to the second exercise, in which we compared a human-written joke with a machine-generated joke. In this case, 70.4% of participants considered the machine-generated output to be a joke. This reinforces our earlier findings that our system is effective at producing content that is widely recognized as humorous. On the other hand, 59.2% of participants identified the human-written humorous explanation as a joke. While this is still a positive result, it suggests that the human-written joke did not resonate as strongly with the audience as the machine-generated one. Additionally, 15.5% of participants selected “None” as their response, meaning they did not consider either the human-written or machine-generated content to be jokes. When breaking down the results by age and education level, the general trend remained consistent—participants were more likely to recognize the machine-generated output as a joke than the human-written one across most demographics. However, one notable exception is the age group of 51 to 65, where the percentage of participants selecting “None” as their answer was higher than in other age ranges. When examining which humorous explanation participants preferred in terms of funniness, we observe an interesting result. Although the machine-generated joke was more frequently recognized as a joke, the gap between the two in terms of funniness is much narrower. Specifically, 43.7% of participants found the machine-generated joke funnier, while 40.9% favored the human-written joke. As before, the remaining participants selected “None” as their response. This near parity in funniness suggests that while the machine-generated joke may have been more readily identified as humor, the human-written joke still held strong appeal in terms of actual comedic impact.

For the final question in this exercise, participants were asked whether they accepted the jokes as an explanation. Compared to the previous exercise, there was an increase in participants selecting “None”, rising to 19.7%. This suggests that a slightly larger portion of the audience did not view either the machine-generated or human-written jokes as valid explanations. Despite this increase, the percentage of participants who accepted the machine-generated and human-written jokes as explanations also increased to 52% and 35.2%, respectively. The increase in both groups—those who accepted the jokes as explanations and those who did not—indicates a more polarized response to the content. On one hand, the majority of participants found the jokes effective enough to serve as explanations, showing that humor can bridge the gap between entertainment and informative content. On the other hand, the grow-

ing number of participants who chose “None” points to the fact that a significant subset of the audience may still struggle with perceiving humor as a legitimate vehicle for conveying factual information. When looking at the demographic breakdown, the age range of 36 to 65 was the most critical in evaluating the jokes as explanations. This group made up 14.1% of the “None” responses, the highest among all age groups. This suggests that middle-aged to older participants may be more demanding when it comes to accepting jokes as valid explanations, possibly preferring more traditional or detailed explanations over humor.

For the final exercise, we compared various forms of creative content—poems, memes, and jokes—to evaluate their strengths and understand participants’ preferences. The first question aimed to determine which format participants found funnier. The results were overwhelmingly one-sided, with 77.5% of participants selecting jokes as the funniest option. This trend was consistent across different age groups and education levels, indicating that humor in the form of jokes tends to resonate broadly with a wide range of audiences. Jokes appear to offer a more universally appealing format for entertainment, regardless of the participant’s background, making them an effective tool for engagement.

Next, we analyzed which creative format participants considered a valid explanation of the information presented. Here, the poem came out on top, with 35.2% of participants accepting it as an explanation. Close behind were jokes, with 33.8% of participants viewing them as explanatory. However, it’s important to note the mixed results between exercises for jokes. This highlights the duality of jokes: they are highly engaging but can struggle to maintain the depth required for conveying detailed or nuanced information. The effectiveness of a joke as an explanation seems to vary depending on its content, style, and delivery. Interestingly, 19.7% of participants selected “None” when asked which format served as an explanation, indicating that a notable segment of the audience did not find any of the creative formats adequate for explaining the information presented. This shows that for certain participants, more traditional, fact-based explanations may still be preferred over creative or humorous approaches, suggesting a limit to the effectiveness of creative formats in educational or factual settings. Finally, memes received the lowest acceptance percentage as an explanation, with only 11.3% of participants considering them informative. This result is not surprising, as memes are often designed for brevity and humor, typically lacking the depth or structure needed to convey detailed information. Memes might be effective at grabbing attention and providing quick commentary, but they fall short when it comes to explaining complex issues, which require more context and elaboration.

After the primary exercises, we included a set of additional questions to better understand participants’ preferences for different types of creative content that could be used to fight misinformation. The first question focused on which type of content participants would be most likely to share, allowing them to select multiple options. This question provides critical insights into which formats participants are more likely to engage with, an important factor in the fight against misinformation. By leveraging the

most shareable content, we can potentially increase the reach of fact-based information. Memes and jokes emerged as the top two formats, with 50.1% and 47.9% of participants selecting them, respectively. This reinforces the notion that humor—whether in the form of memes or jokes—plays a key role in content engagement, making it a valuable tool for spreading factual information. Humor’s ability to quickly capture attention and simplify complex ideas likely contributes to its popularity, as people tend to engage with and share content that is both entertaining and easy to digest.

Interestingly, text explanations also had a strong showing, with 42.3% of participants choosing this option. While text explanations are often seen as less engaging compared to humor, their relatively high pick rate—especially among older participants (aged 36 to 65)—suggests that a significant portion of the audience still values traditional textual explanations. This demographic seems to prefer more thorough, direct methods of communication, which may be perceived as more reliable or trustworthy compared to humor-driven formats. On the lower end of the spectrum, videos and poems were less popular, selected by 25.3% and 12.7% of participants, respectively. The lower preference for videos is somewhat surprising, as video content is often considered highly engaging across social platforms. However, it could suggest that participants find videos too time-consuming or less convenient to share compared to more concise formats like memes and jokes. These results highlight the importance of memes and jokes in online content sharing. Despite memes underperforming in previous parts of the study as explanations, their high shareability rate (50.1%) suggests they are still the go-to format for online engagement. This indicates that memes have significant potential for spreading factual information, provided they are crafted thoughtfully to balance humor with clarity.

The next question aimed to assess participants’ emotional responses when exposed to creative content and whether it contributed to the fight against misinformation. The results were highly encouraging for both our research and the broader objectives of the CIMPLE project. An overwhelming 78.9% of participants reported that creative content made them more open and receptive to the information presented. This positive trend remained consistent across all age groups and education levels, indicating that creative formats like jokes, memes, and other non-traditional methods can break down initial barriers of skepticism or disinterest, fostering a more receptive mindset.

In addition to this, participants were asked whether they believed creative content could simplify complex information in a helpful way. Again, the results were very favorable, with 73.2% agreeing that creative content made complicated ideas more digestible. Only 2.8% felt that it was not helpful, while the remaining participants stated that creative content had no significant impact compared to traditional explanations. These findings highlight the potential of creative approaches to bridge the gap between factual information and audience engagement, simplifying difficult concepts in a way that feels accessible and engaging.

In conclusion, the results demonstrate that creative content not only enhances receptiveness but also

aids in simplifying complex information—key factors in successfully fighting the spread of misinformation.

6.3 Evaluation Discussion

The results of our study highlight the significant potential of machine-generated jokes as a tool to fight misinformation. Throughout the various exercises, the humor generated by our system outperformed or matched human-written content, both in terms of engagement and comedic effectiveness. This is particularly evident when we compare the different sources of humor—our machine-generated jokes were considered funnier and more engaging by a notable majority of participants across multiple questions. Despite one method for joke generation of our system being trained with human-made jokes, we hypothesize that employing two more distinct ways of creating humor managed to give us an edge over the human-written jokes.

These results suggest that while human-created content remains competitive, there is room for improvement. By analyzing the characteristics of human-generated jokes and improving them, we could further refine and enhance the quality of our machine-generated humor.

One of the most promising findings of our research is the fact that machine-generated jokes often passed as human-written, reflecting their high quality and authenticity and on the other hand we have the high percentage in which participants thought human-written jokes were artificially generated, which once again means that there is still room to improve them.

In one exercise, 78.9% of participants reported that creative content (including machine-generated jokes) made them more open and receptive to the information presented, regardless of their age or educational background. Additionally, 73.2% of participants believed that creative content, such as jokes, helped simplify complex information in a useful way. This shows that humor not only engages but can also make complicated ideas more accessible—a crucial element in the fight against misinformation.

Finally, memes and jokes are by far the most shareable content formats, and machine-generated jokes offer a strong combination of engagement and humor. However, human-written content and poems continue to offer greater explanatory power in some cases.

In summary, the results of our study point to machine-generated humor as an effective, high-quality tool for engaging audiences and fighting misinformation. While our system excels at creating engaging, funny content, there are still opportunities to enhance its explanatory power, so we can further refine our system and develop a versatile approach that maximizes both engagement and clarity in the ongoing battle against misinformation.

7

Conclusion

Contents

7.1	Conclusions	91
7.2	Future Work	91

7.1 Conclusions

In this thesis, we transformed jokes into explanatory tools for addressing fake news. To achieve this, we developed three distinct approaches for generating humor. Two of these approaches involved fine-tuning GPT-3.5 Turbo. The first used a custom dataset we created specifically for this task, containing 100 entries of jokes that explained why a particular claim was labeled in a specific way. The second approach employed a larger dataset, consisting of 63,000 jokes sourced from the social media platform Reddit. The third method utilized an algorithmic approach to humor, applying GPT-4 in a prompt-chaining setup to generate the jokes. To ensure the generated jokes remained appropriate, we integrated a moral classifier into the system. This classifier was trained on a dataset of tweets, each paired with their corresponding moral value, to prevent the generation of jokes that could be considered immoral or offensive. Additionally, given the variety of joke-generation methods, we implemented a joke classifier to automate the selection process, ensuring the best joke was chosen from each approach. We evaluated our system through a survey, which revealed that the machine-generated jokes outperformed human-written ones in terms of funniness and engagement. However, there remains room for improvement in explainability, as we observed a trade-off between humor and clarity. Despite this, we concluded that creative content, such as jokes, enhances people's ability to perceive and retain information, making it a valuable tool in the fight against misinformation.

7.2 Future Work

For future work, we aim to enhance and expand our dataset of human-written jokes. As our survey results indicated that these jokes achieved subpar performance compared to the machine-generated ones, improving this dataset could further boost the overall effectiveness of our system. Additionally, we plan to include joke contributions from individuals across a wider range of age groups. This would address the observed bias in our current dataset, where human-written jokes were rated higher by younger participants, likely due to the fact that they were written by people from the same age group. By diversifying the joke creators, we can better appeal to a broader audience. Furthermore, if possible, we would like to collaborate with professional comedians to increase the quality and humor of the jokes.

Another avenue for future exploration is the use of text-to-speech technology to deliver the jokes audibly. We believe that hearing a joke, as opposed to reading it, can significantly enhance its comedic impact. Since auditory delivery often enhances the humor in a joke, integrating text-to-speech could improve the perceived funniness and engagement of the generated content.

Bibliography

- [1] M. Fernandez and H. Alani, "Online misinformation: Challenges and future directions," in *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018, p. 595–602.
- [2] A. Gelfert, "Fake news: A definition," *Informal logic*, vol. 38, no. 1, pp. 84–117, 2018.
- [3] M. Fisher, J. W. Cox, and P. Hermann, "Pizzagate: From rumor, to hashtag, to gunfire in dc," *Washington Post*, vol. 6, pp. 8410–8415, 2016.
- [4] (2021) Cimple - countering creative information manipulation with explainable ai. [Online]. Available: <https://www.chistera.eu/projects/cimple>
- [5] C. S. Taber and M. Lodge, "Motivated skepticism in the evaluation of political beliefs," *American journal of political science*, vol. 50, no. 3, pp. 755–769, 2006.
- [6] S. K. Yeo and M. McKasy, "Emotion and humor as misinformation antidotes," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, 2021.
- [7] P. Quintas and H. S. Pinto, "Report on the state of the art on creative xai," CIMPLE project deliverable 5.1, Tech. Rep., 2022.
- [8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," vol. 4, no. January, 2014, p. 3104 – 3112.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 2017-December, 2017, p. 5999 – 6009.

- [12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [13] Getting started with llm prompt engineering. [Online]. Available: <https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/prompt-engineering>
- [14] Api documentation: Finetuning. [Online]. Available: <https://docs.cohere.com/docs/fine-tuning>
- [15] X. Liu and C. Wang, “An empirical study on hyperparameter optimization for fine-tuning pre-trained language models,” *arXiv preprint arXiv:2106.09204*, 2021.
- [16] M. A. Boden, “Computer models of creativity,” *AI Magazine*, vol. 30, no. 3, pp. 23–23, 2009.
- [17] R. Mihalcea and C. Strapparava, “Computational laughing: Automatic recognition of humorous one-liners,” in *Proceedings of Cognitive Science Conference*, 2005, pp. 1513–1518.
- [18] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in twitter.” in *Proceedings of International Conference on Innovative Computing and Cloud Computing*, 2014, pp. 155–162.
- [19] A. Cattle and X. Ma, “Recognizing humour using word associations and humour anchor extraction,” in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1849–1858.
- [20] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André, “How to shape the humor of a robot-social behavior adaptation based on reinforcement learning,” in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 154–162.
- [21] T. Winters, V. Nys, and D. De Schreye, “Automatic joke generation: Learning humor from examples,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10922 LNCS, p. 360 – 377, 2018.
- [22] A. Valitutti, “How many jokes are really funny? towards a new approach to the evaluation of computational humour generators,” in *Proceedings of 8th International Workshop on Natural Language Processing and Cognitive Science*, 2011, pp. 189–200.
- [23] P. Braslavski, V. Blinov, V. Bolotova, and K. Pertsova, “How to evaluate humorous response generation, seriously?” in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 2018, pp. 225–228.
- [24] T. Winters, “Computers learning humor is no joke,” *Harvard Data Science Review*, vol. 3, no. 2, 2021.

- [25] M. Gervais and D. S. Wilson, "The evolution and functions of laughter and humor: A synthetic approach," *The Quarterly review of biology*, vol. 80, no. 4, pp. 395–430, 2005.
- [26] M. M. Hurley, D. C. Dennett, and R. B. Adams, *Inside jokes: Using humor to reverse-engineer the mind*. MIT press, 2011.
- [27] J. Morreall, Ed., *The Philosophy of Laughter and Humor*. State University of New York Press, 1986.
- [28] E. V. Ezrina and V. Valian, "Do bilinguals get the joke? humor comprehension in mono-and bilinguals," *Bilingualism: Language and Cognition*, vol. 26, no. 1, pp. 95–111, 2023.
- [29] V. Raskin, *Semantic Mechanisms of Humor*, 01 1985, vol. 5.
- [30] R. A. Martin, P. Puhlik-Doris, G. Larsen, J. Gray, and K. Weir, "Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire," *Journal of research in personality*, vol. 37, no. 1, pp. 48–75, 2003.
- [31] W. Ruch, S. Heintz, T. Platt, L. Wagner, and R. T. Proyer, "Broadening humor: Comic styles differentially tap into temperament, character, and ability," *Frontiers in Psychology*, vol. 9, p. 6, 2018.
- [32] S. Heintz and W. Ruch, "From four to nine styles: An update on individual differences in humor," *Personality and Individual Differences*, vol. 141, pp. 7–12, 2019.
- [33] A. Ayçiçeği-Dinn, S. Şişman-Bal, and C. L. Caldwell-Harris, "Are jokes funnier in one's native language?" *Humor*, vol. 31, no. 1, pp. 5–37, 2018.
- [34] K. Binsted, A. Nijholt, O. Stock, C. Strapparava, G. Ritchie, R. Manurung, H. Pain, A. Waller, and D. O'Mara, "Computational humor," *IEEE intelligent systems*, vol. 21, no. 2, pp. 59–69, 2006.
- [35] M. Amin and M. Burghardt, "A survey on approaches to computational humor generation," in *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2020, pp. 29–41.
- [36] K. Binsted and G. Ritchie, "An implemented model of punning riddles," University of Edinburgh, Department of Artificial Intelligence, Tech. Rep., 1994.
- [37] S. Petrović and D. Matthews, "Unsupervised joke generation from big data," in *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2013, pp. 228–232.
- [38] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

- [39] C. Fellbaum, *WordNet: An electronic lexical database*. MIT press, 1998.
- [40] H. Yamane, Y. Mori, and T. Harada, “Humor meets morality: Joke generation based on moral judgement,” *Information Processing & Management*, vol. 58, no. 3, p. 102520, 2021.
- [41] H. D. Schlinger, “Theory of mind: An overview and behavioral perspective,” *The Psychological Record*, vol. 59, pp. 435–448, 2009.
- [42] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaladar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen *et al.*, “Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment,” *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.
- [43] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, “Moral foundations theory: The pragmatic validity of moral pluralism,” in *Advances in experimental social psychology*, 2013, vol. 47, pp. 55–130.
- [44] J. Sjöbergh and K. Araki, “A measure of funniness, applied to finding funny things in wordnet,” in *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, 2009, pp. 236–241.
- [45] L. Burnard, *Users reference guide for the British National Corpus*. Oxford University Computing Services, 1995.
- [46] I. Labutov and H. Lipson, “Humor as circuits in semantic networks,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 150–155.
- [47] W. Ruch, S. Attardo, and V. Raskin, “Toward an empirical verification of the general theory of verbal humor,” *Humor*, vol. 6, no. 2, p. 123 – 136, 1993.
- [48] Z. Yu, J. Tan, and X. Wan, “A neural approach to pun generation,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1650–1660.
- [49] L. Mou, R. Yan, G. Li, L. Zhang, and Z. Jin, “Backward and forward language modeling for constrained sentence generation,” *arXiv preprint arXiv:1512.06612*, 2015.
- [50] K. Church and P. Hanks, “Word association norms, mutual information, and lexicography,” *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.
- [51] H. Ren and Q. Yang, “Neural joke generation,” *Final Project Reports of Course CS224n*, 2017.

- [52] J. Toplyn, "Witscript 3: A hybrid ai system for improvising jokes in a conversation," *arXiv preprint arXiv:2301.02695*, 2023.
- [53] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [54] T. Wu, M. Terry, and C. J. Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–22.
- [55] (2022) Procedural pun and word game generation with tries and gpt-3. [Online]. Available: <https://towardsdatascience.com/procedural-pun-and-word-game-generation-with-tries-and-gpt-3-3661d81a4204>
- [56] A. Valitutti, A. Doucet, J. M. Toivanen, and H. Toivonen, "Computational generation and dissection of lexical replacement humor," *Natural Language Engineering*, vol. 22, no. 5, pp. 727–749, 2016.
- [57] T. Jay, C. Caldwell-Harris, and K. King, "Recalling taboo and nontaboo words," *The American journal of psychology*, vol. 121, no. 1, pp. 83–103, 2008.
- [58] B. He, M. Ahamad, and S. Kumar, "Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2698–2709.
- [59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.
- [60] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.
- [61] A. Santos, "Fake news creative explanations through the use of poetry," Master's thesis, Instituto Superior Técnico, 2022.
- [62] Quadrismegistus, "poesy: Poetic processing, for python". [Online]. Available: <https://github.com/quadrismegistus/poesy>
- [63] I. Annamoradnejad and G. Zoghi, "Colbert: Using bert sentence embedding in parallel neural networks for computational humor," *arXiv preprint arXiv:2004.12765*, 2020.
- [64] L. F. Góes, P. Sawicki, M. Grzes, D. Brown, and M. Volpe, "Is gpt-4 good enough to evaluate jokes?" 2023.

- [65] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, “Sparks of artificial general intelligence: Early experiments with gpt-4,” *arXiv preprint arXiv:2303.12712*, 2023.
- [66] O. Weller and K. Seppi, “The rjokes dataset: a large scale humor collection,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6136–6141.
- [67] Openai fine-tuning guide. [Online]. Available: <https://platform.openai.com/docs/guides/fine-tuning>
- [68] Openai api reference. [Online]. Available: <https://platform.openai.com/docs/api-reference/chat/create#chat-create-seed>
- [69] X developer platform. [Online]. Available: <https://developer.x.com/en/docs/twitter-api/getting-started/make-your-first-request>
- [70] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.
- [71] Jester 1.7m jokes ratings dataset. [Online]. Available: <https://www.kaggle.com/datasets/vikashrajuhaniwal/jester-17m-jokes-ratings-dataset>
- [72] R. Vavekanand and K. Sam, “Llama 3.1: An in-depth analysis of the next-generation large language model,” 2024.



Examples of Jokes Generated by our System

Claim	Label	Joke
Says Joe Biden risks war with Russia because Vladimir Putin doesn't "believe in transgender rights."	Pants-Fire	It's a relief to know that Putin is just being territorial and not transphobic. If he was, he'd be the one getting invaded—Russia has an army, but the LGBTQ+ community has an entire alphabet!
Photo shows Egyptian pyramids being constructed in 2549 BC.	Pants-Fire	Yeah, the photo was taken in 2549 BC. BC stands for basic camera. Just one of those fun facts.
President Trump has come into contact with a potential coronavirus sufferer!	Barely-true	If Trump gets the virus, it will be the first time he has had a positive in a test. However, there is no evidence that President Donald Trump has come into direct contact with someone suffering from coronavirus.
Says Nancy Pelosi delayed coronavirus funding so Dems could campaign on it.	Barely-true	Did you hear the latest rumor about Pelosi? They're saying she delayed the coronavirus funding so "Dems could campaign on it." I guess overseeing the Affordable Care Act wasn't enough excitement! They even suggested she's going to add small business loans to Monopoly's next edition!
The coronavirus was invented and patented in the USA and then planted in China.	False	There is no evidence the coronavirus was "invented", but if it was, it would have been invented in China. After all, they made everything else for us.
Buzz Aldrin admitted that the moon landing "didn't happen."	False	Breaking news: Buzz Aldrin is preparing for his next mission. This one might be even harder than landing on the moon - convincing conspiracy theorists that the whole thing wasn't just a low-budget episode of Star Trek!
Coronavirus Hoax: Fake Virus Pandemic Fabricated to Cover-Up Global Outbreak of 5G Syndrome.	False	I'm tired of this 5G conspiracy theory. I don't know what's worse: the people who believe it, or the people who claim they got 4G Syndrome.
The coronavirus is simply the common cold.	False	If the coronavirus is simply the common cold, then I guess I have been training to fight a pandemic my whole life.
Carbon dioxide is not a problem. The Earth has more than enough land and ocean plant life to metabolize it.	False	"The earth has more than enough land and ocean plant life to metabolize it" - I didn't realize my house plants had to start working overtime!
Switzerland banned mammograms.	False	Switzerland is famous for its neutrality. But I guess it's not neutral on breast cancer screening, with it being available in 14 out of the 26 Swiss cantons.
Donald Trump staged the shooting at his rally in Butler, Pennsylvania.	Pants-Fire	Why is it false that Donald Trump staged the shooting at his own rally in Butler, Pennsylvania? Because he was too busy bleeding from his ear and pumping his fist to show he was okay.
"Every single McDonald's french fry comes exclusively from potatoes grown on Bill Gates-owned farmland."	False	If this was true, they would rebrand the french fries and call them "Microchips". And they would cost \$300 each.
Video shows Elon Musk saying Oreos are "satanic."	Pants-Fire	They say a video shows Elon Musk calling Oreos 'satanic'. But that's a sham, as deceptive as an Oreo promising you'll only eat one! Turns out, the clip was doctored - just like my belief that I could resist a whole packet of Oreos!

Claim	Label	Joke
Congress has officially banned TikTok.	False	TikTok hasn't been banned, but it is being scrutinized by Congress. I guess that's what happens when you give a bunch of old men a platform to watch teenage girls dance: they get weird about it.
"Just Stop Oil activists paint Taylor Swift's private jets."	False	Why did the Just Stop Oil activists paint planes at the airport after Taylor Swift landed, but not her private jet?" Because the players gonna play, play, play, play, play, and the painters gonna paint, paint, paint, paint, paint... just not on Swift's jet!
A jury "ordered Disney and Lucasfilm to pay" actor Gina Carano \$115 million.	False	Gina Carano sued Disney for wrongful termination, and I heard it was partly because they called her a 'Mandalorian' instead of a 'Womandalorian'! The lawsuit is still ongoing, and she hasn't received any money yet, but she must have gotten some Disney+ coupons.
The US has made strides in reducing carbon emissions that other parts of the world have not.	True	America are very good at conserving energy. Must be why they always turn off their brains.
The United States had 589 coal-fired plants 10 years ago, and we're down to 504. We are the only nation that has reduced our reliance (on) coal energy.	Half-true	"Apparently, a decade ago, the US had 589 coal-fired plants while today we're down to 504. Now, that explains the Statue of Liberty turning off her light. She must've finally got her electric bill from the coal-fired power plant."
There will be several thousand dollars in payments going to families with six-figure incomes who've had no income interruption whatsoever.	True	If you're making six figures and still receiving thousands in direct payments from Biden's relief plan, it's like Jeff Bezos getting Amazon Prime for free. Not really necessary but hey, let's give the rich guy the free shipping. And here we thought \$2000 would cover our overdue rent, but for them, it's just another minor stock investment.
Breaking news — US Supreme Court has canceled universal vaccination in the United States.	False	They say the Supreme Court has cancelled universal vaccination in the United States - and you know what, that's as true as my claim of breaking the world record for bench pressing at the gym yesterday!
COVID-19 surges among most vaxxed communities, says Harvard study.	Half-true	It looks like being vaccinated is not enough. I think a new vaccine is in order. It's called 'Common Sense'.
Says Joe Biden has issued more executive orders than anyone in such a short period of time, ever, more than Obama, more than Trump, more than anyone.	True	Biden did sign more executive orders in his first two days, but who can blame him? He was left with a mess bigger than my life.
On key issues including taxes, health care, the Green New Deal, abortion, and guns, there is no daylight between Bernie Sanders and Joe Biden.	Half-true	Bernie Sanders and Joe Biden are starting a cover band called 'Brothers From Another Mother.' They have their differences, like on tax policy, but they still make great music together!
Research illustrates a clear correlation between vitamin D deficiencies and (higher) COVID-19 mortality rates.	True	I guess we can say that he was just trying to shed some light on the situation.

Claim	Label	Joke
Ethylene oxide used in COVID-19 testing swabs literally changes the structure of the building blocks of your own body.	False	If that were true, I would be a building by now, considering how many covid tests I took.



Survey Structure and Content Overview

System evaluation

Misinformation can heavily impact our view of reality, shaping our perspectives on important topics. It often employs eye-catching headlines, altered images, and clever editing to spread more rapidly than accurate information.

It is crucial to uncover and explain these manipulation techniques to help the public understand their purpose and motives. This awareness enables us to make informed decisions based on accurate and reliable information.

To combat misinformation, this survey uses humor as a tool to reveal and address these manipulation techniques. By presenting explanations in a humorous way, we aim to make the process of understanding misinformation both engaging and enlightening.

Your participation will help us explore how humor can be an effective strategy for recognizing and resisting misinformation.

* Indicates required question

Figure B.1: Introduction of the survey.

Form Structure

This form consists of 11 exercises, divided into three different types:

Exercise 1:

- **Overview:**

You will read an explanation and provide feedback on your preferences.

- **Structure:** Each item includes:
 - **Claim:** A statement made by a public person.
 - **Label:** An assessment of the claim's accuracy, categorized as:
 - Pants on Fire
 - False
 - Barely True
 - Half True
 - Mostly True
 - True
 - **Summary Explanation:** An explanation detailing why the claim was assigned the given label.
 - **Explanation:**

A second

humorous explanation for you to evaluate.

Exercise 2:

- **Overview:**

You will read an explanation and provide feedback on your preferences.

- **Structure:** Each item includes:
 - **Claim:** A statement made by a public person.
 - **Explanation:** A humorous

explanation for you to evaluate.

Exercise 3:

- **Overview:**

You will read and compare three different explanations and provide feedback on your preferences.

- **Structure:** Each item includes:
 - **Claim:** A statement made by a public person.
 - **Label:** An assessment of the claim's accuracy, categorized as:
 - Pants on Fire
 - False
 - Barely True
 - Half True
 - Mostly True
 - True
 - **Summary Explanation:** An explanation detailing why the claim was assigned the given label.
 - **Explanations:** Three different explanations for you to evaluate.

Your responses are completely anonymous and will be used solely for research purposes. There are no right or wrong answers, and we encourage you to be honest in your feedback.

Intro

Email *

Your answer

What is your age group ? *

- ☐ < 18
- ☐ 18 - 25
- ☐ 26 - 35
- ☐ 36 - 50
- ☐ 51 - 65
- ☐ 65 - 80
- ☐ 80 >

What is your gender ? *

- ☐ Male
- ☐ Female
- ☐ Non-Binary
- ☐ Prefer not to say
- ☐ Other: _____

What is your education level? *

- ☐ Elementary
- ☐ High-school
- ☐ Bachelor's Degree
- ☐ Master's Degree
- ☐ PhD Degree

On a 1-5 scale (1 being 'uncomfortable but can read in English' and 5 being 'native speaker'), what is your comfort level with the English language? If you do not understand English, please select 0.

- ☐ 0
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5

Jokes

Do you enjoy jokes? *

☐ Yes

☐ No

How would you describe your typical reaction to humorous content? (Select all that apply)

☐ Laughter

☐ Amusement

☐ Disinterest

☐ Skepticism

☐ Other:

What type of jokes do you enjoy the most? (Select all that apply) *

☐ Puns

☐ One-liners

☐ Satire

☐ Dark Humor

☐ Anecdotes

☐ Other: _____

How often do you search for jokes? *

☐ Once a year or less

☐ Once a month or less

☐ Once a week or less

☐ Once a day or less

☐ Multiple times a day

☐ Never

How often do you share jokes with others? *

☐ Once a year or less

☐ Once a month or less

☐ Once a week or less

☐ Once a day or less

☐ Multiple times a day

☐ Never

How often do you tell jokes to others? *

☐ Once a year or less

☐ Once a month or less

☐ Once a week or less

☐ Once a day or less

☐ Multiple times a day

☐ Never

How do you typically find jokes? (Select all that apply) *

☐ Social media (e.g., Reddit, Instagram, Twitter)

☐ Websites dedicated to jokes

☐ Messaging apps (e.g., WhatsApp, Telegram, Messenger)

☐ Word of mouth

☐ Books

☐ Tv Series / Movies

☐ Stand Up Comedy Show/Bar

☐ Podcasts

☐ Live performances or events

☐ Radio

☐ Youtube or other video streaming platforms

☐ Other: _____

Have you ever created a joke that made your friends or family laugh? *

☐ Never

☐ Once

☐ Very rarely but more than once

☐ A few times

☐ Usually

☐ Very often

☐ Almost everyday

☐ Everyday

In your opinion, can jokes be used to effectively convey information or explain concepts?

☐ Yes

☐ No

☐ Sometimes

☐ Unsure

Figure B.2: Humor Preference Questions.

Exercise 1

Claim:

CERN scientists admit they “accidentally” ended the world in 2012 while performing experiments into the Higgs boson particle.

Label: False

Summary Explanation:

The story originated on a satire site.

Experiments in 2012 with a particle accelerator led to the discovery of the Higgs boson particle, and earned CERN researchers a Nobel Prize.

Read the following humorous explanation:

CERN scientists did experiments into the Higgs boson particle in 2012, but they didn't destroy the world. Yet, when my ex called me, I wished they did.

Do you consider this explanation to be a Joke ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Do you like it ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Do you think it is funny ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

How likely would you be to read the full article after seeing the joke, compared to if you only had seen the explanation ? *

	1	2	3	4	5	
Not likely at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely likely

How likely are you to share the joke compared to sharing the explanation? *

	1	2	3	4	5	
Much less likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Much more likely

How important is it for you to know the source of the information presented ?

	1	2	3	4	5	
Not important at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely important

Do you consider it explains well ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

To what extent do you think the joke could be offensive ? *

	1	2	3	4	5	
Not offensive at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very offensive

How engaging do you find the joke compared to the explanation itself ? *

	1	2	3	4	5	
Much less engaging	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Much more engaging

Do you think you are more likely to remember the reason as to why the claim was classified with that Label ?

	1	2	3	4	5	
Much less likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Much more likely

Did you understand the joke?

☐ Yes

☐ No

☐ Partially

☐ Other: _____

After reading the joke, to what extent do you feel it tried to manipulate you or the information in any way ?

	1	2	3	4	5	
Not at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely

Do you consider it explains why the Claim was classified with that Label ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Figure B.3: First type of exercise (News article with machine or human-written joke).

Exercise 8

Claim:

A jury "ordered Disney and Lucasfilm to pay" actor Gina Carano \$115 million.

Joke:

Gina Carano sued Disney for wrongful termination, and I heard it was partly because they called her a 'Mandalorian' instead of a 'Womandalorian! The lawsuit is still ongoing, and she hasn't received any money yet, but she must have gotten some Disney+ coupons.

Source: <https://www.politifact.com/factchecks/2024/jul/01/facebook-posts/no-jury-didnt-order-disney-to-pay-gina-carano-115/>

Do you consider this explanation to be a Joke ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Do you like it ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Do you think it is funny ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Did you understand the joke?

☐ Yes

☐ No

☐ Partially

☐ Other: _____

Do you think it is offensive or it could be offensive to someone ?

	1	2	3	4	5	
Definitely No	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely Yes

Which statement is better represented by the joke?

☐ This claim originated on a self-described satire website. "The Mandalorian" actor Gina Carano has sued Disney for wrongful termination, but the lawsuit continues.

☐ The Mandalorian' actor Gina Carano has been awarded \$115 million after winning her lawsuit against Disney.

☐ None of the above

After reading the joke, to what extent do you feel it tried to manipulate you or the information in any way ?

	1	2	3	4	5	
Not at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely

Do you think it is useful to include the source of the fact at the end of the joke ?

☐ Yes

☐ No

☐ It depends on the context

How would you classify this explanation?

	1	2	3	4	5	
Definitely machine generated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Definitely human written

If you believe the joke was created by a machine, what characteristics made you think so?

Your answer _____

Figure B.4: Second type of exercise (News article without explanation and with a machine generated joke).

Exercise 9

Claim:

People who took the jab are technically 'non human species'.

Label: Pants on fire

Explanation:

People don't lose their "human" status if they get a COVID-19 vaccine.

Joke 1:

Why did the conspiracy theorist break up with his girlfriend after she got the COVID-19 vaccine?

Because he claimed she was technically 'non-human species'! But don't worry, she said she's fine being 'alien-ated' from such nonsense. After all, the truth is, getting vaccinated doesn't change your status from human to extraterrestrial. You won't be growing green scales or antennae after the jab. So, ensure you choose facts over fiction, not comedy over science!

Joke 2:

Why do some people falsely say that those who took the jab are technically 'non-human species'? Because they think they're not Homo sapiens anymore; they're Homo vaccination!

Joke 3:

I took the vaccine and now i'm a freaking unicorn. Unbelievable...

Which of these do you consider to be jokes? (Select all that apply) *

- ☐ First Joke
- ☐ Second Joke
- ☐ Third Joke
- ☐ None

Which one do you like the most ? *

- ☐ First Joke
- ☐ Second Joke
- ☐ Third Joke
- ☐ None

Which one do you think is funnier ?

- ☐ First Joke
- ☐ Second Joke
- ☐ Third Joke
- ☐ None

Which one would you accept as an explanation ?

- ☐ First Joke
- ☐ Second Joke
- ☐ Third Joke
- ☐ None

Which one do you think is more creative ?

- ☐ First Joke
- ☐ Second Joke
- ☐ Third Joke
- ☐ None

Figure B.5: Comparison exercise between: Our system; Domain-Specific Fine-tuned Model and Default GPT-4.

Exercise 10

Claim:
Corona PCR 'test' is implanting a microchip.

Label: Pants on Fire

Explanation:
There is no evidence to support the claim that PCR tests for COVID-19 implant a microchip into the people being tested.
The woman making the claim was unable to explain her fabricated theory, and she has also falsely claimed that COVID-19 testing is a secret blood harvesting operation.

Joke 1:
Seems like another conspiracy theory... If I had a microchip implanted every time I was tested, I would have a much better memory.

Joke 2:
My dad and I have been complaining about this theme for the past few years. Not the microchips, how dumb people are these days. There is no evidence to support this claim.

Which of these do you consider to be jokes? (Select all that apply) *

☐ First Joke

☐ Second Joke

☐ None

Which one do you like the most ? *

☐ First Joke

☐ Second Joke

☐ None

Which one do you think is funnier ?

☐ First Joke

☐ Second Joke

☐ None

Which one would you accept as an explanation ?

☐ First Joke

☐ Second Joke

☐ None

Which one do you think is more creative ?

☐ First Joke

☐ Second Joke

☐ None

How would you classify the explanation for Joke 1?

12345

Definitely machine generated☐☐☐☐☐Definitely human written

If you believe the joke was created by a machine, what characteristics made you think so?

Your answer

How would you classify the explanation for Joke 2?

12345

Definitely machine generated☐☐☐☐☐Definitely human written

If you believe the joke was created by a machine, what characteristics made you think so?

Your answer

Figure B.6: Comparison exercise between: Machine and Human written joke.

109

Claim: COVID-19 “surges among most vaxxed communities, says Harvard study.”

Label: Half-True

Explanation:

A study led by a Harvard researcher found high rates of infection in communities with high rates of vaccination.

It concluded that prevention measures such as handwashing, safe distancing and testing should be employed along with vaccinations to fight the epidemic.

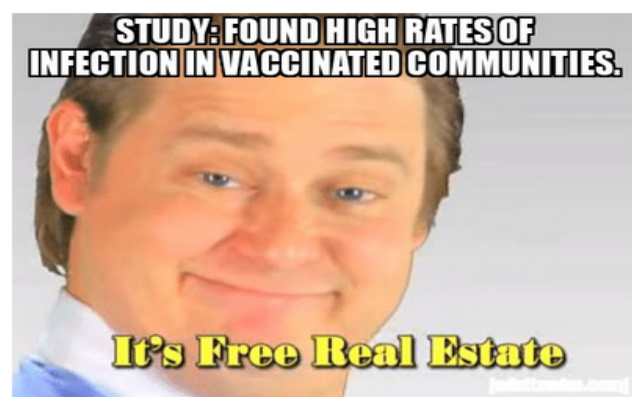
Poem :

No needles, no covid,
you should wash your hands
and vaccinate yourself too.

Joke :

It looks like being vaccinated is not enough. I think a new vaccine is in order.
It's called 'Common Sense'.

Meme:



Which one do you like the most ?

- ☐ Meme
- ☐ None
- ☐ Explanation
- ☐ Joke
- ☐ Poem

Which one do you think is funnier ?

- ☐ Explanation
- ☐ Meme
- ☐ Poem
- ☐ None
- ☐ Joke

Which one would you accept as an explanation ?

- ☐ Joke
- ☐ Meme
- ☐ Poem
- ☐ None

Which one do you think is more creative ?

- ☐ None
- ☐ Meme
- ☐ Joke
- ☐ Poem

Aside from the joke, poem, and meme already shown. If you encountered this claim online, which method would you prefer to read or engage with for clarification? (Select all that apply)

- ☐ Joke
- ☐ Meme
- ☐ Poem
- ☐ Video
- ☐ Text explanation
- ☐ Other: _____

How does encountering Humor or Creative content (jokes, memes, poems) affect your emotional response to the information ?

- ☐ It makes me more receptive and open
- ☐ It makes me skeptical and less likely to engage
- ☐ It has no significant effect on my response
- ☐ Other: _____

Do you think that creative content (jokes, memes, poems, etc.) enhances or detracts from the seriousness of the misinformation issue ?

- ☐ Greatly enhances
- ☐ Somewhat enhances
- ☐ Neither enhances nor detracts
- ☐ Somewhat detracts
- ☐ Greatly detracts

Do you think there are topics too serious to joke about?

- ☐ Yes
- ☐ No
- ☐ Not Sure

If you choose "Yes" in the previous question, please specify which topics you consider too serious to joke about:

Your answer

Do you think creative content (jokes, memes, poems, etc.) simplifies complex information in a way that is helpful or misleading?

- | | | | | | | |
|-----------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| | 1 | 2 | 3 | 4 | 5 | |
| Very misleading | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | Very helpful |

Figure B.7: Comparison exercise between: Jokes, Memes and Poems.

Claim	Label	Joke
Electromagnetic radiation from wireless technologies, including and especially 5G, can cause the same symptoms as COVID-19!	False	"I'm a 5G installation engineer and people are constantly accusing me with bizarre conspiracy theories, such as how 5G is giving them headaches, or that it causes Covid. I think they are completely crazy. 4G must've fried their brains."
Novak Djokovic is the first professional athlete to be banned from a major sporting event for not taking drugs.	False	Australia's COVID-19 policies prevented Djokovic from attending, but at least we had an exciting Australia Open instead of just another Djokovic win.
The Space Foundation has erased the honors previously bestowed on Yuri Gagarin, the first man to ever be in space. His name was stripped 'in light of current events.'	Barely True	The Space Foundation was just trying to protect his legacy from current political tensions. They still think his contributions were out of this world.
Burger King is "closing its doors for good."	False	Before realizing this was fake, I asked McDonald's what they thought, and they said, 'I'm lovin' it!'
In UCLA hundreds of students chanting: 'Israel Israel you can't hide, we want Jewish genocide.'	False	Jews were like "Ah shit, here we go again....". Good thing the protesters were chanting "We charge you with genocide" instead.
Queen Elizabeth II was a reptile.	Pants-Fire	Sure, living that long might make you wonder, but Queen Elizabeth II was definitely human—just like the rest of us.
The International Space Station is in a pool, not in space.	Pants-Fire	If the International Space Station was in a pool, instead of training astronauts for 4 years we could just hire Michael Phelps as a scuba diver.
People who took the jab are technically 'non human species.'	Pants-Fire	That's true, people who took the vaccine have ascended to a new form: Human who can survive Covid.
HarperCollins "removed 64,575 words from the Bible."	False	Imagine if he only left 12.5% of the Bible; that would make us all eighth-theists. Thank god none of this happened and some passages were only relocated.
Says wearing face masks is more harmful to your health than going without one.	False	There's no evidence that wearing masks is harmful to the public, and it's not that difficult to use one. Plus, if you wear glasses with your mask, you might just be entitled to condensation!
The war in Ukraine is "scripted and staged."	Pants-Fire	Testimonies have proven that the war is very real. Vladimir Putin even started calling Ukraine 'Mykraine,' because he can't seem to let it go.

Table B.1: Human Jokes used on our survey.

Claim	Label	Joke
CERN scientists admit they “accidentally” ended the world in 2012 while performing experiments into the Higgs boson particle.	False	CERN scientists did experiments into the Higgs boson particle in 2012, but they didn't destroy the world. Yet, when my ex called me, I wished they did.
Bill Gates is responsible for July 19 global tech outage.	False	Why did Bill Gates get blamed for the July 19 global tech outage? Because people thought he left the Windows open, but it was actually CrowdStrike who forgot to lock the door.
The US has made strides in reducing carbon emissions that other parts of the world have not.	True	America are very good at conserving energy. Must be why they always turn off their brains.
Just Stop Oil activists paint Taylor Swift's private jets.	False	Why did the Just Stop Oil activists paint planes at the airport after Taylor Swift landed, but not her private jet? Because the players gonna play, play, play, play, play, and the painters gonna paint, paint, paint, paint, paint... just not on Swift's jet!
A jury “ordered Disney and Lucasfilm to pay” actor Gina Carano \$115 million.	False	Gina Carano sued Disney for wrongful termination, and I heard it was partly because they called her a 'Mandalorian' instead of a 'Womandalorian'! The lawsuit is still ongoing, and she hasn't received any money yet, but she must have gotten some Disney+ coupons.
Says Joe Biden risks war with Russia because Vladimir Putin doesn't “believe in transgender rights.”	Pants-Fire	It's a relief to know that Putin is just being territorial and not transphobic. If he was, he'd be the one getting invaded—Russia has an army, but the LGBTQ+ community has an entire alphabet!
Breaking news — US Supreme Court has canceled universal vaccination in the United States.	False	So many antivaxxers would lose their minds if this was true, they'd probably throw a party with a dress code of 'No masks!'
President Trump has come into contact with a potential coronavirus sufferer!	False	If Trump gets the virus, it will be the first time he has had a positive in a test. However, there is no evidence that President Donald Trump has come into direct contact with someone suffering from coronavirus.
“Every single McDonald's french fry comes exclusively from potatoes grown on Bill Gates-owned farmland.”	False	If this was true, they would rebrand the french fries and would call them “Microchips”. And they would cost \$300 each. But as it's only one farm that is owned by Bill Gates, we will not be seeing this happen.
The coronavirus was invented and patented in the USA and then planted in China.	False	There is no evidence the coronavirus was “invented,” but if it was, it would have been invented in China. After all, they made everything else for us.
Congress has officially banned TikTok.	False	TikTok hasn't been banned, but it is being scrutinized by Congress. I guess that's what happens when you give a bunch of old men a platform to watch teenage girls dance: they get weird about it.
Jeffrey Epstein found living in New Mexico!	False	I guess it's time to start making 'Weekend at Epstein's' memes.

Claim	Label	Joke
Says wearing face masks is more harmful to your health than going without one.	False	You may not like wearing masks, but there is no evidence that standard masks are harmful. If you don't like wearing masks, wait until you hear about ventilators.
The war in Ukraine is "scripted and staged."	Pants-Fire	News reports, testimonies of citizens-turned-fighters, and a swift global response all fly in the face of baseless claims that Russia's war on Ukraine is "staged." But in Russia everything is staged: the elections, the economy, the media...

Table B.2: Machine Generated Jokes used on our survey.