A Two-Model Approach for Humour Style Recognition

Mary Ogbuka Kenneth[1], Foaad Khosmood[2], Abbas Edalat[1]
[1]Algorithmic Human Development group, Department of Computing, Imperial College London, UK
[2]Computer Engineering Department, California Polytechnic State University, USA
m.kenneth22@imperial.ac.uk, foaad@calpoly.edu, a.edalat@imperial.ac.uk

Abstract

Humour, a fundamental aspect of human communication, manifests itself in various styles that significantly impact social interactions and mental health. Recognising different humour styles poses challenges due to the lack of established datasets and machine learning (ML) models. To address this gap, we present a new text dataset for humour style recognition, comprising 1463 instances across four styles (self-enhancing, self-deprecating, affiliative, and aggressive) and non-humorous text, with lengths ranging from 4 to 229 words. Our research employs various computational methods, including classic machine learning classifiers, text embedding models, and DistilBERT, to establish baseline performance. Additionally, we propose a two-model approach to enhance humour style recognition, particularly in distinguishing between affiliative and aggressive styles. Our method demonstrates an 11.61% improvement in f1-score for affiliative humour classification, with consistent improvements in the 14 models tested. Our findings contribute to the computational analysis of humour in text, offering new tools for studying humour in literature, social media, and other textual sources.

1 Introduction

Humour recognition is a multidimensional task influenced by various theories and manifested through diverse styles. There are various humour theories, such as relief, incongruity, and superiority theories (Morreall, 2011, 2012; Scheel and Gockel, 2017). The relief theory highlights the role of humour in relaxation, while the incongruity theory suggests that we find something funny when we notice a mismatch or contradiction between what we expect in a situation and what actually happens. The superiority theory suggests that people may laugh at other people's misfortunes in an effort to demonstrate their superiority.

These theories not only explain why we find things humorous but also why we laugh as a response. In recent decades, evolutionary psychology has introduced a new perspective on laughter itself, known as the *play* theory (Martin and Ford, 2018): laughter developed as a play signal in higher primates in their mock fights to indicate non-aggressive intent.

Laughter, therefore, is more than just a reaction to humour; it serves various functions, including promoting mental, emotional, and physical well-being. This idea forms the basis for laughter therapy, a cognitive-behavioural treatment designed to induce laughter and reduce stress, tension, anxiety, and sadness (Yim, 2016). However, as Martin et al. (2003) noted, not all humour is beneficial—some forms can even harm relationships with others or oneself.

Considering its impact on well-being, Martin et al. (2003) categorised humour into four styles: self-enhancing, self-deprecating, affiliative, and aggressive. Affiliative and self-enhancing humour are beneficial to psychological well-being. Affiliative humour fosters social bonding, while self-enhancing humour involves maintaining a positive outlook without harming oneself or others, often employed as a coping mechanism in difficult situations (Edalat, 2023; Kenneth et al., 2024; Hampes, 2007; Plessen et al., 2020). In contrast, aggressive and self-deprecating humour can be harmful. Aggressive humour, rooted in superiority theory, belittles or mocks others, whereas self-deprecating humour seeks approval by making oneself the target of jokes (Khramtsova and Chuykova, 2016; Kuiper et al., 2016; Veselka et al., 2010).

In artificial intelligence (AI), humour is considered AI-complete (Shani et al., 2021; Strapparava et al., 2011; Kenneth et al., 2024), meaning that a system capable of producing and recognising human-like humour would possess general intelligence. Despite the importance of humour, most computational efforts have focused on laughter de-

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 259–274
November 16, 2024. ©2024 Association for Computational Linguistics

tection (Vargas-Quiros et al., 2023; Matsuda and Arimoto, 2023; Inoue et al., 2022), classification (Tanaka and Campbell, 2014) and generation (Inoue et al., 2022), as well as humour detection (Oliveira et al., 2020; Jaiswal et al., 2019; Chauhan et al.), and humour generation (Luo et al., 2019; He et al., 2019; Yu et al., 2018), with little emphasis on humour styles and their links to well-being. Kenneth et al. (2024) identified a gap in the current ML landscape: the lack of datasets and models specifically designed to recognise these four humour styles.

Building on the gaps identified by Kenneth et al. (2024), this study addresses the lack of an established dataset and ML models for recognising the four humour styles: self-enhancing, self-deprecating, affiliative, and aggressive. We draw on Martin et al. (2003), who defined and validated these styles, providing the theoretical basis for our classification task. Additionally, Edalat (2023)'s work on self-initiated humour protocols (SIHP) informs how different humour styles can enhance well-being, while Amjad and Dasti (2022) research on the link between humour styles, emotion regulation, and subjective well-being highlights the potential applications of our work in psychological and clinical contexts. By integrating these insights, we aim to develop a comprehensive approach to humour style recognition grounded in psychological theory and applicable to real-world scenarios. The key contributions of this paper are:

1. Introduction of a new text dataset for humour style recognition, addressing the lack of established datasets. This dataset is publicly available to the community.
2. Baseline evaluations using various ML classifiers and models.
3. Development of a two-model approach for improved humour style recognition.
4. Extensive evaluation of the proposed two-model approach.

2 Related Works

Humour recognition and classification are active research areas in NLP and multi-modal analysis. While our focus is on humour style recognition, we draw insights from related fields like general humour detection and sarcasm detection.

Weller and Seppi (2020) compiled a dataset of 550,000 jokes from Reddit posts, using user ratings and engagement metrics as quantifiable humour quality measurements. However, the dataset's reliance on Reddit data alone may introduce biases and limit generalisability. Our study addresses this by introducing a more diverse dataset specifically tailored for humour style recognition from various online sources.

Oliveira et al. (2020) explored humour recognition in Portuguese text, achieving a 75% f1-score using Naive Bayes, Support Vector Machine, and Random Forest classifiers. However, their work was limited to binary classification of headlines and one-liners. Our approach extends this by focusing on multi-class classification of humour styles in both short and long texts.

Tang et al. (2022) created a dataset and classification model for sub-types of inappropriate humour, using large language models like BERT. While relevant, their focus on inappropriate humour differs from our goal of recognising humour styles linked to psychological well-being.

Kamal and Abulaish (2020) targeted self-deprecating humour, one of the four styles we examine. Their use of specific feature categories (self-deprecating pattern, and word embedding) informs our feature engineering process. However, our study broadens the scope to include all four humour styles.

Christ et al. (2022a,b) developed models for humour recognition in German football press conferences. Although their work yielded promising results, it was limited to the MuSe humour challenge and the Passau-SFCH German dataset, unlike our broader approach.

Sarcasm detection is closely related to humour style recognition since it is often used in aggressive and self-deprecating humour styles. Liang et al. (2021) used an interactive graph convolution network for multi-modal sarcasm detection, highlighting the importance of contextual cues. This technique could be adapted to distinguish humour styles.

Jinks (2023) improved sarcasm detection with a two-step fine-tuning process using RoBERTa, a method that could enhance humour style classification given the subtle differences between styles.

Fang et al. (2024) introduces the Single-Stage Extensive Semantic Fusion model for multi-modal sarcasm detection by concurrently processing and fusing multi-modal inputs in a unified framework. This approach could be adapted for humour style recognition, when we expand our dataset to include multi-modal features in the future.

Although these studies contribute to the detection of humour and sarcasm, there is a gap in recognising the four humour styles defined by Martin et al. (2003). Our work fills this gap by creating a dedicated dataset and developing classification models tailored to these humour styles.

3 Dataset Collection and Annotation

A significant challenge in identifying humour styles automatically is the lack of annotated datasets suitable for training machine learning models. To address this, we created a comprehensive dataset comprising 1,463 instances from various sources:

1. 983 jokes from several well-known websites where jokes were labelled by users or editors.
2. 280 non-humorous text instances from the ColBERT dataset (Annamoradnejad and Zoghi, 2020).
3. 200 instances from the Short Text Corpus [1], consisting of 150 jokes and 50 non-jokes

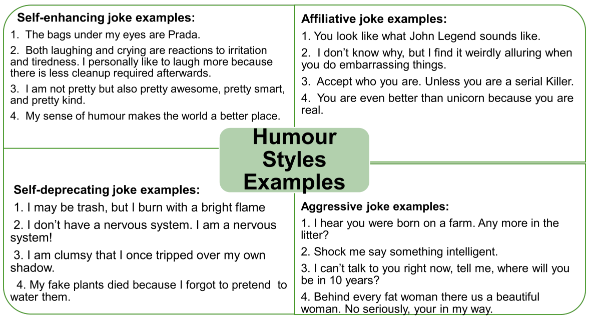
After annotation, the dataset consists of 298 instances of self-enhancing humour, 265 of self-deprecating humour, 250 of affiliative humour, 318 of aggressive humour, and 332 neutral instances, with text lengths ranging from 4 to 229 words. This distribution ensures balanced representation across the different humour styles and neutral text.

3.1 Data Sources and Labelling

The 983 jokes were extracted from sources like Reader's Digest, Parade, Bored Panda, Laugh Factory, Pun Me, Independent, Cracked, Reddit, Tastefully Offensive and BuzzFeed. We labelled each joke based on the original labels, definitions, or tags given on the websites, mapping them to our categories based on humour theory. Table 1 summarises these mappings, illustrating how the website tags correspond to our humour style labels.

Equivalence Classes (Website Keywords)	Humour Styles
Dark (inappropriate) Jokes	Aggressive
Insult	Aggressive
Icebreakers Jokes for Work Meetings	Affiliative
International Day of Happiness	Affiliative
Friendship	Affiliative
Family jokes	Affiliative
Classroom jokes	Affiliative
Self-deprecating	Self-deprecating
Self-love	Self-enhancing
Self-care	Self-enhancing

Table 1: Terminological Equivalence Classes



Self-enhancing joke examples:
1. The bags under my eyes are Prada.
2. Both laughing and crying are reactions to irritation and tiredness. I personally like to laugh more because there is less cleanup required afterwards.
3. I am not pretty but also pretty awesome, pretty smart, and pretty kind.
4. My sense of humour makes the world a better place.

Affiliative joke examples:
1. You look like what John Legend sounds like.
2. I don't know why, but I find it weirdly alluring when you do embarrassing things.
3. Accept who you are. Unless you are a serial Killer.
4. You are even better than unicorn because you are real.

Self-deprecating joke examples:
1. I may be trash, but I burn with a bright flame
2. I don't have a nervous system. I am a nervous system!
3. I am clumsy that I once tripped over my own shadow.
4. My fake plants died because I forgot to pretend to water them.

Aggressive joke examples:
1. I hear you were born on a farm. Any more in the litter?
2. Shock me say something intelligent.
3. I can't talk to you right now, tell me, where will you be in 10 years?
4. Behind every fat woman there us a beautiful woman. No seriously, your in my way.

Figure 1: Joke Examples for Each Humour Style

For example, in Table 1 the "Dark (inappropriate)" tag was mapped to the aggressive style because dark or inappropriate jokes are identified as being cruel, morbid, or offensive to some, which aligns with the characteristics of aggressive humour (Tang et al., 2022). Further details on these mappings are available in Appendix B.

To simulate real-life scenarios where users might input non-humorous text, we added 280 non-humorous instances from the ColBERT dataset (Annamoradnejad and Zoghi, 2020), labelled as Neutral.

Figure 1 presents random examples from the dataset for each humour style. Additionally, word clouds showing the most common words associated with each humour style in the created dataset are provided in Appendix C.

3.2 Dataset Composition and Potential Biases

Each humour style in our dataset was primarily sourced from different websites (see Table 11 in Appendix A for details). The use of diverse websites, catering to various audiences and content styles, helps mitigate biases that could arise from relying on a single source. However, since the jokes were collected in English, there may be language biases, as humour often involves nuances and idioms specific to certain languages and cultures.

By aggregating data from multiple websites, we aimed to reduce inherent biases from any single source and provide comprehensive coverage of different humour styles, enhancing the robustness of the dataset. However, most websites (except Reader's Digest and Laugh Factory) featured jokes from only one humour type, potentially introducing idiosyncratic styles that could lead the classifier to learn spurious correlations.

To address this concern and further diversify our dataset, we included an additional 200 jokes from

[1] Short Text Corpus (`https://github.com/CrowdTruth/Short-Text-Corpus-For-Humor-Detection`)

the existing Short Text Corpus joke dataset[1] and have them annotated by six human annotators. Details of the Short Text Corpus[1] and the annotation process are discussed further in the following subsection.

3.3 Annotation Process and Inter-annotator Agreement

Building on our efforts to address potential biases in our dataset composition, we took additional steps to ensure the robustness of our data. To mitigate potential biases from idiosyncratic styles of the individual websites, we randomly selected 200 instances from the Short Text Corpus[1], dividing them into two sets of 100 samples. This corpus was chosen for its diversity, featuring both short and long jokes from more than seven sources, as well as non-jokes from three sources. In contrast, the ColBERT dataset (Annamoradnejad and Zoghi, 2020) was not used here because it consists solely of Reddit jokes, which would not address the issue of spurious correlations.

To further ensure the reliability of our annotations, we recruited six Ph.D. candidates from Africa, Asia, and Europe to serve as annotators, bringing a diverse range of analytical perspectives to the task. Each set of 100 samples was independently annotated by three annotators, who were provided with humour style definitions and asked to classify each instance as self-enhancing, self-deprecating, aggressive, affiliative, or neutral. A majority vote determined the final label for each instance.

Fleiss' Kappa was used to assess inter-annotator agreement. The results showed fair agreement levels:

1. First 100 samples: Fleiss' Kappa = 0.2651
2. Second 100 samples: Fleiss' Kappa = 0.2841

Despite the relatively low Kappa values, further analysis showed substantial agreement among at least two annotators:

1. For the first set of 100 samples: 91 samples had at least two annotators agreeing on the label and 9 instances had all three annotators disagreeing.
2. For the second set of 100 samples: 95 samples had at least two annotators agreeing on the label and 5 instances had all three annotators disagreeing.

To resolve the 14 instances (9 in the first set, 5 in the second) where all three annotators disagreed, indicating no majority vote, we used four Large Language Models (LLMs) chatbots: ChatGPT, Claude, Microsoft Copilot, and HuggingChat - to classify the jokes. We prompted the LLMs to categorise each joke instance as self-enhancing, self-deprecating, aggressive, affiliative, or neutral. Each of the 14 instances then had seven labels (from the 4 LLMs and 3 human annotators), and the majority label was assigned. Table 12 in Appendix D provides examples of instances where annotators disagreed, along with the annotators' and LLMs' labels.

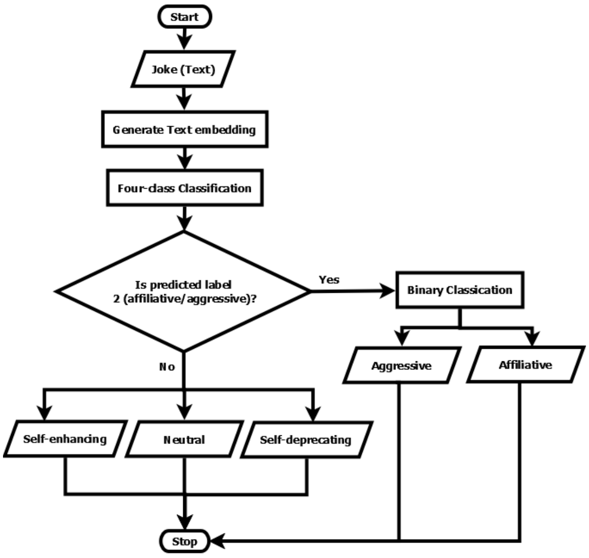


Figure 2: Flowchart illustrating the proposed Two-Model Approach for Humour Style Recognition

These disagreements highlight the subjective nature of humour interpretation, which can be influenced by cultural differences, personal experiences, and individual preferences (Lu, 2023). This subjectivity is a natural aspect of humour annotation, and our use of multiple annotators and LLMs helps to mitigate its impact.

4 Methodology

This study employs two different approaches for humour style recognition: the single-model and the two-model approach. A total of 14 models were evaluated, including Naive Bayes, Random Forest, XGBoost (each with six different text embeddings), and DistilBERT. Figure 2 illustrates the two-model approach, which first classifies humour instances into broader groups before refining to specific styles.

4.1 Classifiers and Embedding Models

4.1.1 Classifiers

The selection of classifiers was based on their suitability for the task at hand and efficiency in low-resource settings, avoiding resource-intensive large language models such as GPT4 and LLaMA prone to overfitting on small datasets due to their complex architectures (Schur and Groenjes, 2024; Diwakar and Raj, 2024; Berfu B et al., 2020):

Naive Bayes (NB): A probabilistic classifier based on the Bayes Theorem, assuming conditional independence of features given the target class (Berrar, 2019).

Random Forest (RF): A bagging ensemble classifier using majority voting from multiple decision trees (Jin, 2020).

eXtreme Gradient Boosting (XGBoost): A boosting ensemble classifier aggregating predictions of several weak learners, with regularisation to prevent overfitting (Jiang et al., 2019).

DistilBERT: A condensed BERT variant, offering faster performance and memory efficiency while maintaining competitive performance on NLP tasks (Sanh et al., 2019).

4.1.2 Sentence Embedding Models

To capture distinct linguistic nuances and improve classification performance, we selected six embedding models from the top 20 on the Massive Text Embedding Benchmark (MTEB) leaderboard. These models were chosen for their robustness, efficiency, speed, and lightweight memory usage:

- General Text Embeddings (GTE) and GTE Upgraded (ALI) (Li et al., 2023)
- BAAI General Embedding (BGE) (Xiao et al., 2022; Zhang et al., 2023)
- Matryoshka Representation Learning and Binary Quantization (MRL) (Lee et al., 2024)
- Universal AnglE Embedding (UAE) (Li and Li, 2023)
- Multilingual E5 Text Embeddings (MUL) (Wang et al., 2024)

These embeddings were combined with RF and XGBoost classifiers for humour style recognition.

4.2 Single-Model Approach

In this approach, a single ML model is trained to classify the input text into one of the five classes: self-enhancing (**label 0**), self-deprecating (**label 1**), affiliative (**label 2**), aggressive (**label 3**), and neutral (**label 4**). This approach treats the task as a multi-class classification problem, where the model needs to distinguish between all five classes simultaneously.

To provide insight into the single-model performance, Figure 3 presents the confusion matrices for the 5-fold cross-validation results of four models: Naive Bayes (NB) 3a, GTE+RF 3b, MUL+XGBoost 3c, and UAE+RF 3d.

4.3 Two-Model Approach

To address limitations observed in the single-model approach, particularly in distinguishing affiliative humour, we developed a two-model approach. This method, inspired by previous studies (Khan et al., 2022; Van Lam et al., 2011; Demidova, 2021), improves classification performance by breaking down the problem into multiple steps.

The rationale behind this approach is to first separate the instances into broader groups and then focus on the more challenging task of distinguishing between affiliative and aggressive humour styles. This strategy is informed by an analysis of misclassified samples from the cross-validation and test set evaluation of the single-model approach, which revealed that affiliative humour was predominantly misclassified as aggressive humour. This pattern of misclassification is clearly illustrated in the cross-validation confusion matrices shown in Figure 3.

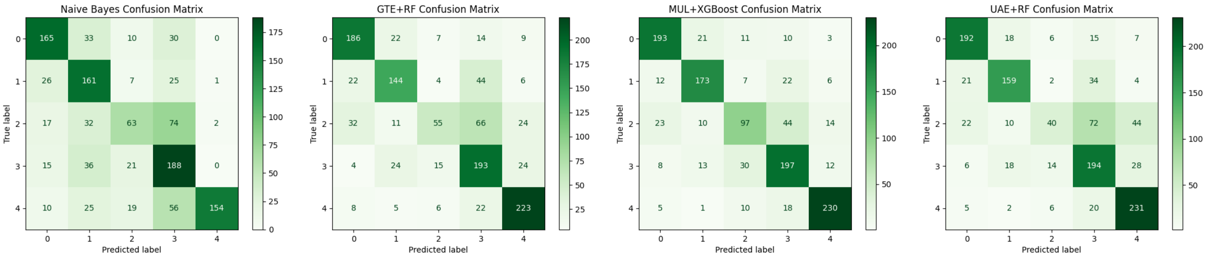
The two-model approach involves two sequential steps:

1. **Step 1: Four-Class Classification Model:** Train an ML model to distinguish between self-enhancing, self-deprecating, neutral, and a combined affiliative/aggressive class.
2. **Step 2: Binary Classification Model:** Train a separate binary classification model to distinguish between affiliative and aggressive instances from the combined class in step 1.

This approach allows for optimising overall performance by combining the best-performing models for each subtask.

4.4 Experimental Setup

The humour styles dataset was split 80/20 for training and testing, randomised using a fixed seed of 100 to ensure reproducibility. We used 5-fold cross-validation for all experiments to validate model performance and prevent overfitting. For the NB classifier, we used a smoothing parameter of 1. The RF and XGBoost classifiers were implemented using their default hyperparameters. The DistilBERT



(a) Naive Bayes (b) GTE+RF (c) MUL+XGBoost (d) UAE+RF

Figure 3: 5-Fold Cross Validation Confusion Matrix

model was fine-tuned for 5 epochs with a weight decay of 0.01, warmup steps of 500, and a training batch size of 8, using the default learning rate scheduler provided by the Hugging Face Transformers library.

4.5 Evaluation Metrics

Model performances were evaluated using standard metrics: accuracy, precision, recall, and f1-score. Accuracy measures overall performance, precision quantifies the ratio of true positives to predicted positives, recall assesses the model's ability to identify actual positives, and f1-score represents the harmonic mean of precision and recall. Furthermore, the Wilcoxon signed-rank test was used to compare the single-model and two-model approaches, determining the statistical significance of the performance differences between these approaches.

5 Results and Discussions

Experiments for the single-model and two-model approaches were conducted on Fourteen models: NB, RF + six embedding models, XGBoost + six embedding models and DistilBERT.

5.1 Baseline Model (Single-Model Approach)

Tables 2 and 3 show the mean accuracy and macro-mean f1-score of the 5-fold cross-validation for different models and embedding techniques, respectively. The results highlight the robustness and generalisability of our models across different data splits.

Table 4 presents the overall performance for the five-class classification. MUL+RF, ALI+RF, and DistilBERT performed best with accuracies and f1-scores of 77.1% and 76.6%, 77.8% and 77.3%, and 75.4% and 75.2%, respectively.

While the single-model approach achieved decent overall performance, Table 5 reveals that all models struggle to identify affiliative humour accurately. Despite high overall accuracy, this approach fails to differentiate affiliative humour from other styles, particularly aggressive humour, as shown in Figure 3, highlighting a critical issue.

This misclassification may stem from affiliative humour sometimes containing slightly aggressive components, as noted by Martin et al. (2003). For example: ***JOKE:*** *'To be happy with a man, you must understand him a lot and love him a little. To be happy with a woman, you must love her a lot and not try to understand her at all'.* (**LABEL:** *True:'Affiliative'*, *Predicted: 'Aggressive'*)

This joke attempts to playfully highlight gender differences, aiming for camaraderie. However, its misclassification as aggressive likely stems from the presence of gender stereotypes that could be misconstrued as demeaning. This example illustrates how subtle nuances in tone, context, and intent can lead to misclassifications between affiliative and aggressive humour.

5.2 Two-Model Approach

To address the challenge of misclassifying affiliative humour as aggressive, we implemented a two-model approach, consisting of a four-class model and a binary-class model. The performance of these individual models is presented in Tables 6 and 7, which show their accuracy and macro-mean f1-score, respectively. Among the four-class models, MUL+XGBoost achieved the highest performance, with an accuracy of 85.3% and a macro-mean f1-score of 85.1%. In contrast, the binary-class model ALI+XGBoost outperformed the other models, with an accuracy and f1-score of 80.0%.

The results of the two-model approach, which combines the four-class and binary models, are presented in Tables 8 and 9. This approach yields improved overall performance compared to the single-model method, with the best results

Model	**NB (%)**	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
		BGE	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Five-Class	62.5	69.2	68.5	69.7	67.0	70.4	71.9	69.7	71.2	72.1	71.3	73.0	76.1	75.9
Four-Class	66.0	75.4	74.3	74.3	72.8	76.5	79.1	78.3	78.2	79.8	79.1	78.8	82.1	82.4
Binary-Class	74.8	73.9	78.8	75.9	74.8	77.2	78.1	71.9	79.5	74.1	75.2	76.6	80.3	78.3

Table 2: Mean Accuracy of 5-Fold Cross-Validation for the Various Classification Models

Model	**NB (%)**	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
		BGE	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Five-Class	61.4	65.2	65.9	65.9	63.5	67.9	69.0	67.7	70.1	71.0	70.1	71.6	74.9	74.6
Four-Class	63.7	73.1	72.5	73.5	71.66	75.3	77.9	77.2	77.9	79.6	78.7	78.2	82.0	81.9
Binary-Class	74.1	71.4	77.1	73.8	72.9	75.6	76.1	70.0	78.8	73.1	73.8	75.4	79.5	77.7

Table 3: Macro-mean F1-Score of 5-Fold Cross-Validation for the Various Classification Models

	NB (%)	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
		BGE	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Precision	64.1	72.7	71.4	76.5	65.0	72.9	72.7	73.6	70.1	73.7	68.5	**77.6**	76.8	75.6
Recall	62.5	70.3	71.6	74.3	64.0	72.7	72.1	74.0	70.6	72.7	68.4	**77.6**	77.4	75.1
F1-score	61.4	68.5	69.2	72.6	61.7	71.8	70.8	72.6	69.7	72.3	67.6	**77.3**	76.6	75.2
Accuracy	61.8	70.3	71.7	74.4	64.5	73.0	72.7	73.7	71.3	73.0	68.9	**77.8**	77.1	75.4

Table 4: Performance of the Single-Model Approach

	NB (%)	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT(%)**
Humour Styles		**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Self-enhancing	61.7	80.3	81.9	82.8	70.7	76.9	85.0	80.3	81.6	80.0	73.2	82.6	**86.2**	79.4
Self-deprecating	66.0	72.5	76.7	**80.5**	65.9	70.5	66.7	77.1	67.4	75.9	71.3	79.1	77.6	76.7
Affiliative	39.2	40.5	34.9	46.5	33.7	54.5	47.3	50.0	48.5	57.4	48.0	**64.9**	63.0	60.2
Aggressive	56.4	62.7	69.1	72.0	58.9	71.3	67.6	67.2	65.6	66.7	62.8	**74.8**	67.7	70.8
Neutral	83.6	86.3	83.4	81.3	79.5	85.7	87.1	88.2	85.3	81.6	82.6	85.1	**88.7**	**88.7**

Table 5: Macro-mean F1-score for each humour style for the Single-Model Approach

achieved by the combination of MUL+XGBoost and ALI+XGBoost, which attained a f1-score of 78.0% and an accuracy of 77.8%. Notably, in Tables 8 and 9, MUL+XGBoost was consistently used as the four-class model in combination with various binary models (embeddings + RF or XGBoost), as it had previously demonstrated the best performance among the four-class models.

The Wilcoxon signed-rank test results (Table 10) statistically validate the improvements observed in the two-model approach. Significant improvements (p-value < 0.05) are evident for most metrics and humour styles, except aggressive humour (p-value = 0.1189). The two-model approach consistently outperforms the single-model approach, with average increases ranging from 3.42% to 4.91% across precision, recall, f1-score, and accuracy.

Notably, the two-model approach significantly improved affiliative humour classification, with an 11.61% increase in f1-score. All 14 models showed improvement for affiliative humour under this approach, suggesting more robust and accurate classification, especially for previously challenging categories like affiliative humour.

The cross-validation results (Tables 2 and 3) further support the robustness of our findings. The five-class models' cross-validation accuracies and macro-mean f1-scores generally align with final test set accuracies and macro-mean f1-scores, indicating good generalisation. The four-class and binary-class models achieved even closer alignment, suggesting robust generalisation.

In summary, the two-model approach demonstrates superior performance in humour style recognition, particularly in identifying affiliative humour, with improved performance and generalisability across various metrics.

6 Conclusion

Automatic recognition of humour styles is a valuable yet challenging task with significant implications for digital humanities research, particularly in areas such as mental health, content moderation, and social media discourse. This study addresses the lack of established resources by introducing a new dataset of 1,463 instances across four humour styles and non-humour, while providing baseline evaluations of various models.

	NB (%)	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
Models		**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Four-Class	73.0	76.5	80.5	77.1	75.1	80.9	83.6	80.5	80.9	81.2	76.8	82.6	**85.3**	82.6
Binary-Class	76.7	70.0	74.2	74.2	73.3	75.8	78.3	71.7	71.7	74.2	70.8	**80.0**	78.3	79.2

Table 6: Performance Accuracy of Four-Class and Binary-Class Individual Models

	NB (%)	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
Models		**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Four-Class	70.5	73.2	78.4	76.1	73.5	80.1	82.4	79.4	78.8	80.3	75.7	81.3	**85.1**	81.8
Binary-Class	76.3	69.5	73.9	73.8	73.1	75.8	78.3	71.6	71.4	73.9	70.7	**80.0**	78.3	79.2

Table 7: Macro-mean F1-score of Four-Class and Binary-Class Individual Models

Four-Class Model ->	**NB**	**MUL + XGBoost**						**MUL + XGBoost**						**DistilBERT**
Binary-Class Model ->	**NB (%)**	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
		BGE	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Precision	72.7	75.0	77.9	78.1	78.3	78.6	78.5	75.5	76.3	77.2	76.5	**78.6**	78.2	76.8
Recall	67.2	73.6	76.3	76.3	76.3	76.9	77.2	74.5	74.9	75.9	74.9	**77.8**	77.5	74.8
F1-score	67.4	73.5	76.3	76.2	76.3	77.1	77.4	74.8	75.0	75.9	75.1	**78.0**	77.7	75.3
Accuracy	67.6	73.4	76.1	76.1	76.1	76.8	77.1	74.4	74.7	75.8	74.7	**77.8**	77.5	75.4

Table 8: Performance of the Two-Model Approach

Four-Class Model ->	**NB**	**MUL + XGBoost**						**MUL + XGBoost**						**DistilBERT**
Binary-Class Model ->	**NB (%)**	**Random Forest (%)**						**XGBoost (%)**						**DistilBERT (%)**
		BGE	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	**BGE**	**GTE**	**UAE**	**MRL**	**ALI**	**MUL**	
Self-enhancing	56.8	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	86.4	80.3
Self-deprecating	66.7	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	80.9	75.6
Affiliative	67.6	50.5	59.8	59.0	61.0	66.7	65.5	57.1	56.4	58.2	58.2	**66.1**	63.9	61.2
Aggressive	64.0	61.5	65.7	66.2	64.8	63.3	65.7	61.1	62.9	65.7	61.4	68.2	68.8	**71.2**
Neutral	81.7	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.4	88.1

Table 9: Macro-mean F1-score for each humour style for the Two-Model Approach

	Precision	**Recall**	**F1-Score**	**Accuracy**	**Self-enhancing**	**Self-deprecating**	**Affiliative**	**Aggressive**	**Neutral**
Wilcoxon Statistics	0.0	3.0	0.0	0.0	8.0	3.0	0.0	27.0	10.0
P-value	0.000122	0.000610	0.000122	0.00220	0.0031	0.0006	0.0001	0.1189	0.0052
Average (Single-Model)	72.23	71.66	70.52	71.83	78.76	73.14	49.19	66.69	84.79
Average (Two-Model)	77.01	75.29	75.43	75.25	83.85	79.51	60.80	65.04	87.90
Model Difference	4.79	3.63	4.91	3.42	5.09	6.37	11.61	-1.65	3.11
# of improved models out of 14	14	13	14	12	13	13	14	4	11

Table 10: Wilcoxon Sign-Rank Test to Compare the Single-model and Two-model Approaches

The dataset and research have significant implications in three key areas:

1. **Mental Health**: Automatically identifying humour styles can enhance mental health research by enabling large-scale analysis of social media content. Different humour styles may correlate with various mental health indicators, potentially aiding in early detection of conditions such as depression or anxiety. For example, frequent use of self-deprecating humour might signal underlying mental health concerns.
2. **Content Moderation**: The dataset can contribute to more refined content moderation systems on social media platforms. By distinguishing between different humour styles, moderators can better identify potentially harmful content disguised as humour, such as aggressive or self-defeating jokes, while allowing for benign forms of humour that enhance online interactions.
3. **Social Media Discourse**: Automatic recognition of humour styles can provide valuable insights into social dynamics and communication patterns across various online communities. This can help researchers understand how different humour styles influence online discussions, shape public opinion, and contribute to the spread of information or misinformation.

Our initial single-model approach struggled to accurately recognise affiliative humour, with f1-scores ranging from 39.2% to 64.9%. To address this, we developed a two-model approach consisting of a four-class model (merging affiliative and aggressive styles) followed by a binary model distinguishing between these styles. Extensive evaluation demonstrated the effectiveness of this approach in improving affiliative humour recognition, achieving f1-scores of 50.5% to 66.1%, while maintaining good performance for other styles. Furthermore, this approach offers flexibility in combining the best models for each sub-task, optimising overall performance.

By introducing this dataset and baseline evaluations, we aim to catalyse further research and development in these critical areas of digital humanities, ultimately enhancing our understanding of humour and its multifaceted impact on human communication.

7 Dataset Availability

The dataset and models implemented in this study are available to the community via the link in the footnote [2]. Additionally, thirty instances from the dataset are included in Appendix E.

8 Limitations and Future Works

This study has several limitations. The dataset, consisting of 1,463 instances, is relatively small, which may limit the model's generalisation capabilities. Additionally, the inherent subjectivity of humour, along with the observed inter-rater agreement and annotation disagreements, underscores the challenges in consistently labelling humorous content. The focus on English-centric jokes may also introduce biases and language-specific nuances.

Future research could focus on collecting larger and more diverse datasets from various languages and sources to improve the robustness of the model. Leveraging transfer learning methods, such as intermediate fine-tuning on pre-trained language models, could enhance performance, especially when data is limited. Exploring multimodal approaches that incorporate visual, auditory, and contextual cues, as well as personalised models that adapt to individual preferences, could provide deeper insights into humour styles. Furthermore, investigating generative models for producing humorous content in specific styles presents a promising direction for further exploration.

Despite these limitations, this study lays the groundwork for humour style recognition, paving the way for extensive future research on computational humour analysis and its applications in digital humanities.

9 Acknowledgement

This research was supported by the Petroleum Technology Development Fund (PTDF) of Nigeria.

References

Arooba Amjad and Rabia Dasti. 2022. Humor styles, emotion regulation and subjective well-being in young adults. *Current Psychology*, 41(9):6326–6335.

Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using BERT Sentence Embedding in Parallel Neural Networks for Computational Humor.

Berfu B, Ali H, Arzucan¨özgür Arzucan¨, and Arzucan¨ozg Arzucan¨ozg¨ Arzucan¨özgür. 2020. Analyzing ELMo and DistilBERT on Socio-political News Classification. In *Language Resources and Evaluation Conference*, pages 11–16.

Daniel Berrar. 2019. Bayes ' Theorem and Naive Bayes Classifier Bayes ' Theorem and Naive Bayes Classifier. *Encyclopedia of Bioinfor- matics and Computational Biology*, (January 2018):0–18.

Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. All-in-One: A Deep Attentive Multitask Learning Framework for Humour, Sarcasm, Offensive, Motivation, and Sentiment on Memes. Technical report.

Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva Maria Meßner, Andreas König, Alan Cowen, Erik Cambria, and Björn W. Schuller. 2022a. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. In *MuSe 2022 - Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge*, pages 5–14. Association for Computing Machinery, Inc.

Lukas Christ, Shahin Amiriparian, Alexander Kathan, Niklas Müller, Andreas König, and Björn W. Schuller. 2022b. Multimodal Prediction of Spontaneous Humour: A Novel Dataset and First Results. *Transactions on Affective Computing*, 20(10):1–15.

Cleveland Clinic. 2024. Self-Love: Why It's Important and What You Can Do To Love Yourself.

Silvie Cooper and David Dickinson. 2013. Just jokes! Icebreakers, innuendo, teasing and talking: The role of humour in HIV/AIDS peer education among university students. *African Journal of AIDS Research*, 12(4):229–238.

Ron Deiter. 2000. The Use of Humor as a Teaching Tool in the College Classroom. Technical Report 2.

[2] Humour Styles Dataset: `https://github.com/MaryKenneth/Two_Model_Humour_Style`

Liliya A. Demidova. 2021. Two-stage hybrid data classifiers based on svm and knn algorithms. *Symmetry*, 13(4).

Diwakar and Deepa Raj. 2024. DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions. In *Microbial Data Intelligence and Computational Techniques for Sustainable Computing*, pages 93–106.

Abbas Edalat. 2023. Self-initiated humour protocols: An algorithmic approach for learning to laugh. *PsyArXiv*, 5:1–14.

Hong Fang, Dahao Liang, and Weiyu Xiang. 2024. Single-Stage Extensive Semantic Fusion for multimodal sarcasm detection. *Array*, 22:100344.

William Kodom Gyasi. 2023. Humor as an ice breaker in marital tension: A family communication perspective. *Mediterranean Journal of Social & Behavioral Research*, 7(2):103–111.

William P Hampes. 2007. The Relation Between Humor Styles and Empathy. *Europe's Journal of Psychology*, 6(3):34–45.

He He, Nanyun Peng, and Percy Liang. 2019. Pun Generation with Surprise. In *Proceedings of NAACL-HLT*, pages 1734–1744, Minneapolis. Association for Computational Linguistics.

Lena Hedin, Ingrid Höjer, and Elinor Brunnberg. 2012. Jokes and routines make everyday life a good life-on 'doing family' for young people in foster care in Sweden. *European Journal of Social Work*, 15(5):613–628.

Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9.

Arunima Jaiswal, Monika Anshu, Mathur Prachi, and Sheena Mattu. 2019. Automatic Humour Detection in Tweets using Soft Computing Paradigms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)*, pages 172–176.

Daniela Jeder. 2015. Implications of Using Humor in the Classroom. *Procedia - Social and Behavioral Sciences*, 180:828–833.

Yu Jiang, Guoxiang Tong, Henan Yin, and Naixue Xiong. 2019. A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters. *IEEE Access*, 7:118310–118321.

Wei Jin. 2020. Research on Machine Learning and Its Algorithms and Development. *Journal of Physics: Conference Series*, 1544:1–6.

Jarrad Jinks. 2023. Intermediate Task Ensembling for Sarcasm Detection. In *Bramer, M., Stahl, F. (eds) Artificial Intelligence XL*, volume 14381, pages 19–32. Springer, Charm, Oldeenburg.

Ashraf Kamal and Muhammad Abulaish. 2020. Self-deprecating Humor Detection: A Machine Learning Approach. In Le-Minh Nguyen, Xuan-Hieu Phan, Kôiti Hasida, and Satoshi Tojo, editors, *Computer Lingustics*, volume 1215 of *Communications in Computer and Information Science*, pages 483–484. Springer Singapore, Singapore.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. Systematic Literature Review: Computational Approaches for Humour Style Classification. Technical report.

Muhammad Umar Khan, Sumair Aziz, Khushbakht Iqtidar, Galila Faisal Zaher, Shareefa Alghamdi, and Munazza Gull. 2022. A two-stage classification model integrating feature fusion for coronary artery disease detection and classification. *Multimedia Tools and Applications*, 81(10):13661–13690.

I. I. Khramtsova and T. S. Chuykova. 2016. Mindfulness and self-compassion as predictors of humor styles in US and Russia. *Social Psychology and Society*, 7(2):93–108.

Nicholas Kuiper, Gillian Kirsh, and Nadia Maiolino. 2016. Identity and Intimacy Development, Humor Styles, and Psychological Well-Being. *Identity*, 16(2):115–125.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open Source Strikes Bread - New Fluffy Embeddings Model (emb2024mxbai).

Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. *arXiv preprint*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. *arXiv preprint*.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs. In *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, pages 4707–4715. Association for Computing Machinery, Inc.

Jackson G. Lu. 2023. Cultural differences in humor: A systematic review and critique.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative Adversarial Network for Pun Generation.

Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*, 2nd edition. Academic Press.

Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, 37:48–75.

Takuto Matsuda and Yoshiko Arimoto. 2023. Detection of laughter and screaming using the attention and CTC models. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023-August, pages 1025–1029. International Speech Communication Association.

John Morreall. 2011. *Comic relief: A comprehensive philosophy of humor*. John Wiley & Sons, Ltd.

John Morreall. 2012. Philosophy of Humor.

National Institute of Mental Health. 2024. Caring for Your Mental Health.

Hugo Gonçalo Oliveira, André Clemêncio, and Ana Alves. 2020. Corpora and Baselines for Humour Recognition in Portuguese. In *Proceedings ofthe 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 1278–1285, Marseille. European Language Resources Association.

Constantin Y. Plessen, Fabian R. Franken, Christoph Ster, Rebecca R. Schmid, Christoph Wolfmayr, Anna Maria Mayer, Marc Sobisch, Maximilian Kathofer, Katrin Rattner, Elona Kotlyar, Rory J. Maierwieser, and Ulrich S. Tran. 2020. Humor styles and personality: A systematic review and meta-analysis on the relations between humor styles and the Big Five personality traits. *Personality and Individual Differences*, 154.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS*.

Tabea Scheel and Christine Gockel. 2017. *Humor at Work in Teams, Leadership, Negotiations, Learning and Health*, volume 1.

Amir Schur and Sam Groenjes. 2024. Comparative Analysis for Open-Source Large Language Models. In *Communications in Computer and Information Science*, volume 1958 CCIS, pages 48–54. Springer Science and Business Media Deutschland GmbH.

Chen Shani, Nadav Borenstein, and Dafna Shahaf. 2021. How Did This Get Funded?! Automatically Identifying Quirky Scientific Achievements. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 14–28. Association for Computational Linguistics.

Carlo Strapparava, Oliviero Stock, and Rada Mihalcea. 2011. Computational Humour. In *Emotion-Oriented systems*, pages 609–634.

Hiroki Tanaka and Nick Campbell. 2014. Classification of social laughter in natural conversational speech. *Computer Speech and Language*, 28(1):314–325.

Leonard Tang, Alexander Cai, Steve Li, and Jason Wang. 2022. The Naughtyformer: A Transformer Understands Offensive Humor. *axXiv*.

Le Van Lam, Ian Welch, Xiaoying Gao, and Peter Komisarczuk. 2011. Two-stage classification model to detect malicious web pages. In *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, pages 113–120.

Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. 2023. Impact of Annotation Modality on Label Quality and Model Performance in the Automatic Assessment of Laughter In-the-Wild. *IEEE Transactions on Affective Computing*.

Livia Veselka, Julie Aitken Schermer, Rod A. Martin, and Philip A. Vernon. 2010. Relations between humor styles and the Dark Triad traits of personality. *Personality and Individual Differences*, 48(6):772–774.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*.

Orion Weller and Kevin Seppi. 2020. The rJokes Dataset: a Large Scale Humor Collection. Technical report.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. *arXiv preprint*.

Jong Eun Yim. 2016. Therapeutic benefits of laughter in mental health: A theoretical review.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1650–1660. Association for Computational Linguistics.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve Anything To Augment Large Language Models. *arXiv preprint*.

A Humour Style Websites

The website sources for the different humour styles, along with their corresponding links, are listed in Table 11.

Humour Styles	Website
Aggressive	Parade Laugh factory Reader's digest Pun Me
Affiliative	Reader's digest Independent Happy Numbers Laugh factory Team building
Self-Deprecating	Tastefully Offensive Bored Pandas Cracked Reddit Buzz Feed
Self-Enhancing	Put the Kettle On Silk and Sonder Carley Schweet Joyful through it all Laura Conteuse

Table 11: List of websites from which jokes were taken.

B Mapping Jokes to Humour Style Labels

Although certain humour websites from which the jokes were extracted do not explicitly categorise the humour as "aggressive," "affiliative," or "self-enhancing," there are reasonable justifications for associating the humour found on those sites with the respective humour styles, based on the content and intended audience. This section outlines the keywords and rationale for mapping jokes to the original labels, definitions, or tags provided for jokes on the websites.

B.1 Aggressive Humour

Aggressive humour is characterised by jokes, insults, or humorous remarks that are intended to disparage, belittle, or target particular individuals or groups. This type of humour often involves sarcasm, mockery, and put-downs, and it can be perceived as offensive or hostile by the targeted parties.

B.1.1 Equivalence classes (Website Keywords)

- **Dark (inappropriate) Jokes:** Dark (inappropriate) jokes are identified as being cruel, morbid, or offensive to some, which aligns with the characteristics of aggressive humour (Tang et al., 2022).
- **Insult:** Insult is an offensive remark or action intended to mock or belittle the target (Cambrige Dictionary (`https://rb.gy/l0b2sz`)). Insult is a key characteristic of aggressive humour (Martin et al., 2003).

B.2 Affiliative Humour

Affiliative humour is characterised by jokes, witty remarks, or humorous anecdotes that are intended to amuse others, facilitate social interactions, and strengthen relationships. This type of humour is non-hostile, benign, and often used to create a positive, inclusive atmosphere.

B.2.1 Equivalence classes (Website Keywords)

- **Icebreakers Jokes for Work Meetings:** This jokes are typically used to create a relaxed and friendly environment in professional or group settings.They are meant to facilitate social interactions and put people at ease, which aligns with the goals of affiliative humour (Cooper and Dickinson, 2013).
- **International Day of Happiness:** Jokes shared on occasions like the International Day of Happiness are typically intended to spread positivity, joy, and laughter among people. Such jokes are designed to bring people together and create a shared experience of amusement, which aligns with the goals of affiliative humour
- **Friendship:** Jokes meant to be shared among friends are often used to strengthen bonds, create shared laughter experiences, and reinforce the positive aspects of friendship. This type of humour is non-threatening and aimed at building connections, which is a characteristic of affiliative humour.
- **Family jokes:** Jokes shared within families are often intended to create a sense of bonding, shared laughter, and enjoyment. Family jokes are generally non-offensive and serve to strengthen familial relationships, which is a characteristic of affiliative humour (Hedin et al., 2012; Gyasi, 2023).
- **Classroom:** Humour shared between teachers and students, or within educational settings, is often used to create a positive and engaging learning environment. These jokes are likely meant to connect with students and foster a sense of camaraderie, which is in line with affiliative humour (Deiter, 2000; Jeder, 2015)

B.3 Self-deprecating Humour

Self-deprecating humour is a type of humour in which individuals make fun of their own flaws, weaknesses, or mistakes. It involves mocking or belittling oneself in a humorous way.

B.3.1 Equivalence classes (Website Keywords)

- **Self-deprecating:** The title on the websites directly states that the quotes and captions are "self-deprecating," implying that they involve humour directed at oneself in a self-mocking or self-effacing manner. Given the explicit use of the term "self-deprecating" in the titles of the websites, the jokes found on these sites are labelled as self-deprecating jokes.

B.4 Self-enhancing Humour

Self-enhancing humour is characterised by jokes, witty remarks, or humorous anecdotes that are intended to promote a positive self-image, boost self-confidence, and enhance one's sense of self-worth. This type of humour often involves self-affirmation, playful boasting, or exaggerating one's positive qualities in a light-hearted and non-hostile manner.

B.4.1 Equivalence classes (Website Keywords)

- **Self-love:** Self-love, involves deliberately prioritising oneself, supporting your needs and desires, and respecting your limitations. It entails refraining from self-criticism, regret, shame, or guilt, and confronting uncomfortable emotions rather than avoiding them (Cleveland Clinic, 2024). Self-love is closely tied to self-enhancement, as it involves promoting a positive self-image and boosting self-confidence. Humorous texts that encourage self-love can be seen as self-enhancing, as they aim to make one feel better about themselves and promote self-acceptance.
- **Self-care:** Self-care is the intentional practice of dedicating time to activities that promote overall well-being, encompassing both physical and mental health benefits. By effectively managing stress, reducing the risk of illness, and increasing energy levels, self-care fosters a healthier lifestyle (National Institute of Mental Health, 2024). A key aspect of self-care is cultivating a positive self-image and nurturing one's own well-being. In this context, humorous texts that aim to promote self-affirmation and boost self-confidence can be seen as self-enhancing, as they seek to enhance an individual's self-worth and overall sense of well-being.

C Word Clouds of Humour Styles Phrases

Figure 1 provides a selection of examples from the dataset for each humour style. Figures 4, 5, 6, and 7 represent word clouds of the most common words associated with each of the humour styles in the created dataset. Figures 4 and 6 reveal a prevalence of positive phrases, including self-love, laughter, good, love, friends, and happy, in self-enhancing and affiliative humour. In contrast, Figures 5 and 7 highlight the presence of negative phrases, including ugly, fat, stupid, bad, depression, and mistakes, in self-deprecating and aggressive humour styles.

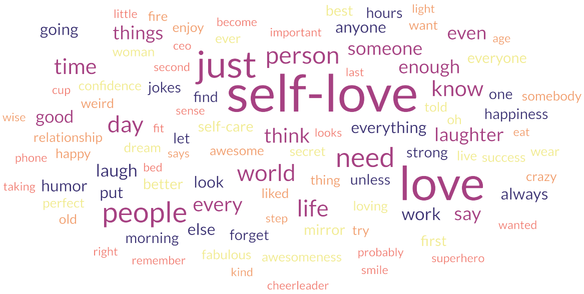


Figure 4: Most Frequent Self-Enhancing Phrases

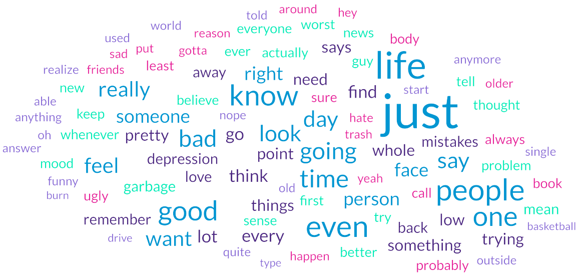


Figure 5: Most Frequent Self-Deprecating Phrases

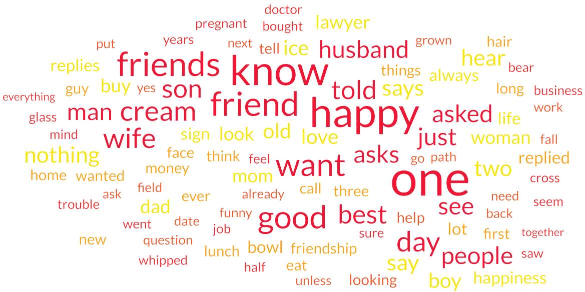


Figure 6: Most Frequent Affiliative Humour Phrases

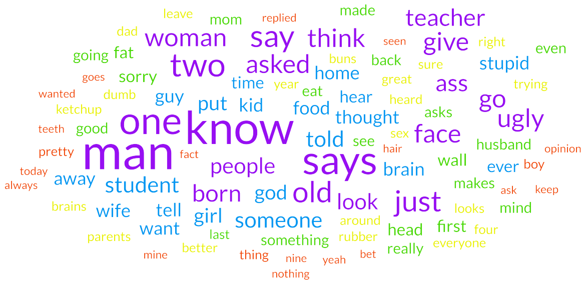


Figure 7: Most Frequent Aggressive Humour Phrases

D Annotation Disagreement

Table 12 presents examples of jokes where the three human annotators (A1, A2, and A3) for each of the jokes disagreed on the annotation labels, as discussed in Section 3.3. The labels are interpreted as follows: self-enhancing (0), self-deprecating (1), affiliative (2), aggressive (3), and neutral (4). To gain insight into the annotation process, each rater, along with the LLM models, was asked to provide a rationale for their label assignments. Below, we summarise the reasons behind the label assignments for four of these jokes.

JOKE: *Insanity is hereditary, - You get it from your children.*

The annotators had varying interpretations of this joke, which are summarised below:

- **Affiliative -2 (Copilot, ChatGPT, and A3):** This joke is light-hearted and relatable, playing on the common experiences of parenting. It is inclusive and bonding, fostering a sense of shared understanding.
- **Self-enhancing -0 (A1 and Claude):** The joke-teller uses a playful and light-hearted tone to poke fun at themselves, without being overly self-critical. The joke does not come across as aggressive or hostile towards anyone.
- **Self-deprecating -1 (HuggingChat):** Although not aggressive or mocking, the joke can be seen as self-deprecating. It humorously comments on the challenges of parenting, implying that the joke-teller is not immune to the stresses of parenthood.
- **Aggressive -3 (A2):** In contrast, one annotator interpreted the joke as aggressive, believing that it mocks and belittles parents.

JOKE: *Don't worry if you're a kleptomaniac, you can always take something for it.*

- **Affiliative -2 (Claude, Copilot, ChatGPT, and A3):** This joke uses a lighthearted and playful tone to make a humorous comment about kleptomania, potentially creating a sense of shared understanding and camaraderie. Its intention is to be humorous rather than offensive.
- **Self-enhancing -0 (A1):** The joke-teller attempts to reframe their mental health disorder in a positive light, presenting it in a humorous and optimistic way.
- **Self-deprecating -1 (HuggingChat):** The joke can be seen as self-deprecating, as it humorously acknowledges the potentially embarrassing or shameful nature of kleptomania.
- **Aggressive -3 (A2):** In contrast, one annotator interpreted the joke as aggressive, believing that it belittles and mocks individuals with kleptomania, a mental health disorder.

JOKE: *Always follow your dreams. Except for that one where you're naked at work.*

- **Affiliative -2 (Copilot, ChatGPT, and A1):** This joke takes a common piece of advice and adds a humorous twist that many people can relate to. It aims to create a sense of shared amusement over a common anxiety.
- **Self-enhancing -0 (A2):** The speaker presents themselves as someone who can laugh at their own imperfections and embarrassing moments, showcasing a positive coping mechanism.
- **Self-deprecating -1 (HuggingChat and Claude):** The joke uses self-deprecation to poke fun at the embarrassing nature of certain dreams, with the joke-teller willingly making themselves the target of the humour.
- **Neutral -4 (A3):** This joke is interpreted as a neutral observational joke, lacking strong emotional undertones and instead focusing on a humorous observation.

JOKE: *Never get stuck behind the Devil in a Post Office queue! The Devil can take many forms.*

- **Neutral -4 (HuggingChat, Copilot, ChatGPT, and A3):** This joke is a lighthearted commentary on the frustrations of waiting in line, without any specific target or malicious intent.
- **Affiliative -2 (Claude and A1):** The joke creates a sense of shared understanding and camaraderie around the common experience of waiting in line, which most people can relate

to.

- **Aggressive (A2):** In contrast, one annotator interpreted the joke as aggressive, as the Devil's representation of negative traits or behaviours could be seen as a critique of people in general.

E Sample Jokes Dataset

In this section, we showcase a random selection of thirty samples from our jokes dataset (see Table 13). Each sample consists of the joke content paired with its corresponding label, providing a glimpse into the dataset's composition and structure. For reference, the labels are interpreted as follows:

- Self-enhancing: 0
- Self-deprecating: 1
- Affiliative: 2
- Aggressive: 3
- Neutral: 4

Jokes	A1	A2	A3	Hugging Chat	Claude	Copilot	Chat-GPT
Insanity is hereditary, - You get it from your children.	0	3	2	1	0	2	2
Gravity doesn't exist: the earth sucks.	0	4	3	4	0	4	4
Did you hear about the Scottish Kamikaze pilot? He crashed his plane in his brother's junkyard	3	4	2	3	3	3	3
Biology grows on you	4	3	2	2	4	4	4
Don't worry if you're a kleptomaniac, you can always take something for it	0	3	2	1	2	2	2
To steal from one is plagiarism. To steal from many is research	2	3	4	1	0	2	2
If all else fails, lower your standards	2	1	4	1	1	1	1
There are only 3 things that tell the truth: 1 - Young Children 2 - Drunks 3 - Leggings	1	3	4	2	4	2	2
Never get stuck behind the Devil in a Post Office queue! The Devil can take many forms.	2	3	4	4	2	4	4
Always follow your dreams. Except for that one where you're naked at work.	2	0	4	1	1	2	2

Table 12: Annotation Disagreement

Jokes	Labels
Is that your nose or are you eating a banana?	3
Q: Why did the witches' team lose the baseball game? A: Their bats flew away.	2
Act your age, not your shoe size.	3
I may be trash, but I burn with a bright flame	1
Yeah, I know. I hate me too.	1
"The secret of staying young is to live honestly, eat slowly, and lie about your age."	0
"I got it all together. But I forgot where I put it."	0
A man on a date wonders if he'll get lucky. A woman already knows.	2
Here's how unfair the tax system is in each state	4
Is a death sentence really a death sentence?	4
Trump's new military plan will cost 150 billion dollars – at the very least	4
He is so short, his hair smells like feet.	3
You should be in commercials for birth control.	3
"The road to success is dotted with many tempting parking spaces."	0
If I had a face like yours, I'd sue my parents!	3
"I'm not perfect, but I'm perfectly me."	0
Why don't scientists trust atoms? Because they make up everything!	2
Don't mind me. I'm just having an existential crisis. Move along, folks.	1
I can't talk to you right now, tell me, where will you be in 10 years?	3
A wise woman once said, "fuck this shit" and lived happily ever after.	0
He is depriving a village somewhere of an idiot.	3
Dad: "Can I see your report card, son?" Son: "I don't have it." Dad: "Why?" Son: "I gave it to my friend. He wanted to scare his parents."	2
"The elevator to success is out of order. You'll have to use the stairs, one step at a time."	0
"I'm not the kind of guy who has a huge weight problem, but I am the kind of guy who could really put the brakes on an orgy. Everyone would be like, 'Was he invited? Why is he eating a cake?' I've never been in an orgy, but I feel like it'd be like what happens when I try to play pickup basketball: No one passes me the ball, and everyone asks me to keep my shirt on."	1
"I'm a self-love junkie. Can't get enough of this good stuff!"	0
"If I could rearrange the alphabet, I'd put 'U' and 'I' together."	2
"Let your light shine bright so the other weirdos can't find you"	0
Did you hear about the magic tractor? It turned into a field.	2
I don't have a nervous system. I am a nervous system!	1
How to build muscle: proven strength lessons from milo of croton	4

Table 13: Samples from the Humour Styles Dataset