King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT362: Principles of Data Science
1st Semester 1447 H

Humor in Headlines: Sociolinguistic Variation.

Section No.	Students Names	Student ID
80657	Basmah Alrashid	442202996
	Noura Alamro	444200941
	Shooq Alawadah	444201083
	Reema Alraqibah	444201069

Supervised by: Dr.Abeer Aldayel

Table of Contents

Final Report

Introduction:

The project explores the variability of humor in headlines through a sociolinguistic lens, recognizing that humor is deeply influenced by social context and language variation. The main research question guiding this work is:

“How does language, context, and social factors influence the perception and structure of humor in short textual content, such as headlines?”.

In Phase 1, the focus was on collecting and curating data from both a structured jokes dataset and user-generated Reddit content from humor-related subreddits. This phase established the foundation for the analysis by gathering a diverse corpus of humorous texts that reflect both curated and spontaneous expressions of humor in online communities. The collected data captured a wide range of linguistic styles, audience interactions, and cultural nuances embedded in humor.

Phase 2 built upon this foundation by emphasizing data processing, cleaning, and exploratory data analysis (EDA). The raw data were carefully inspected to identify and correct issues such as missing values, duplicates, and irregular text formatting. Text normalization and preprocessing techniques were applied to ensure consistency across datasets, while additional features such as emoji usage, punctuation frequency, and title length were extracted to capture sociolinguistic and paralinguistic cues relevant to humor.

Following the cleaning phase, exploratory data analysis was conducted to uncover trends and relationships within the data. The analysis focused on understanding how language and structure influence engagement through patterns in upvotes, comments, word frequency, and emoji use. Comparative analysis between the primary Reddit dataset and the secondary Kaggle jokes dataset revealed both commonalities and distinctions in humor styles, showing how contextual and cultural factors shape the way humor is expressed and perceived across platforms.

Phase 3 builds on the previous stages by moving into the modelling and communication components of the project. In this phase, the prepared dataset is used to develop predictive models that explore how linguistic and structural features contribute to humor classification. This involves selecting an appropriate modelling task, establishing a baseline model, and training at least two additional models using different algorithms. The models are compared using suitable evaluation metrics, and the best-performing approach is chosen based on the project's objectives. Throughout this phase, all modelling steps, decisions, and observations are documented, ensuring a clear and systematic workflow that connects the earlier exploratory insights to the final analytical outcomes.

Phase 1

1. Data Sources

1.1 Hugging Face Short Jokes Datasets

URL: https://huggingface.co/datasets/Fraser/short-jokes
Description: A large dataset containing ~230,000 short jokes in English. It includes brief joke texts without additional metadata.

Observations & Features:

joke (string): The text of the joke
Potential Biases:
Representation: Primarily English-language jokes; may underrepresent non-English-speaking cultural humor.
Measurement: Content is user-submitted or collected from online sources; the humor is subjective and unannotated for context.
Historical Biases: Some jokes may include outdated or culturally insensitive references.

1.2 Reddit Humor Posts

Subreddits: r/funny, r/nottheonion, r/humor
Collection Method: API using PRAW in Python

Data Collected:

title (headline of the post)
score (popularity)
num_comments
created_utc (posting timestamp)

Cleaning Steps:

1. Removed posts with profanity using better profanity
2. Filtered posts with unusual symbols or encoding issues (mojibake)
3. Dropped unnecessary columns (id, url, author, selftext)

Observations: Collected 200 posts per mode (hot, new, top) per subreddit, resulting in a total of ~1,800 raw posts. After cleaning, a subset was retained for analysis.

Potential Biases:

- Representation: Reddit users skew younger and more tech-savvy; not globally representative.
- Measurement: Post titles vary widely in style and length; humor interpretation is subjective.
- Historical Biases: Posts may reflect trends or memes relevant only to the collection timeframe.

2. Objectives

The cleaned dataset will enable the following insights and analyses:

- Examine linguistic features of humor in short text, such as puns, wordplay, and cultural references.
- Analyze the relationship between post popularity (score, comments) and type of humor.
- Explore variability in humor across subreddits and user communities.
- Develop models to classify humor type or predict audience engagement.
- Study sociolinguistic factors that influence humor perception and reception.

3. Method

Collect raw jokes from Hugging Face and Reddit using API and web scraping techniques.
Clean data by removing posts with profanity, weird symbols, or encoding issues.
Retain only essential columns (title, score, num_comments, created_utc) for analysis.
Analyze the dataset to identify humor patterns, linguistic features, and correlations with engagement metrics.

4. Challenges Faced:

Profanity & Sensitive Content: Required automated filtering to ensure clean dataset.
Encoding Issues: Some Reddit posts contained mojibake or unusual symbols, requiring regex filtering.
Bias in Representation: Both sources skewed toward English-language humor; Reddit data biased toward active users and specific cultural references.

API Limitations: Rate limits and post retrieval restrictions required batching and iterative collection.

Phase 2

1. Primary Data

This section presents a detailed analysis of the cleaned Reddit dataset. The aim is to explore patterns in headline length, emoji usage, subreddit distribution, and engagement metrics, and to identify features relevant to humor perception.

1.1 Dataset Overview

- Total posts (after cleaning): 4,508
- Number of subreddits: 6 (funny, jokes, memes, Oneliners, cleanjokes, standupshots)
- Posts containing emojis: 78 (1.7%)
- Dataset shape: 4,508 rows × 7 columns (subreddit, type, title, score, num_comments, emojis, has_emoji)

The dataset consists of short, user-generated headlines with minimal missing data and proper text standardization, making it suitable for linguistic and engagement analysis.

1.2 Text Characteristics

- Average title length: 48.24 characters
- Average word count: 9.22 words
- Average emoji count (posts with emojis): 1.50

Most headlines are concise, reflecting the nature of social media humor. Posts containing emojis are significantly shorter both in characters and words. This suggests that emojis often accompany short, punchy text for expressive effect.

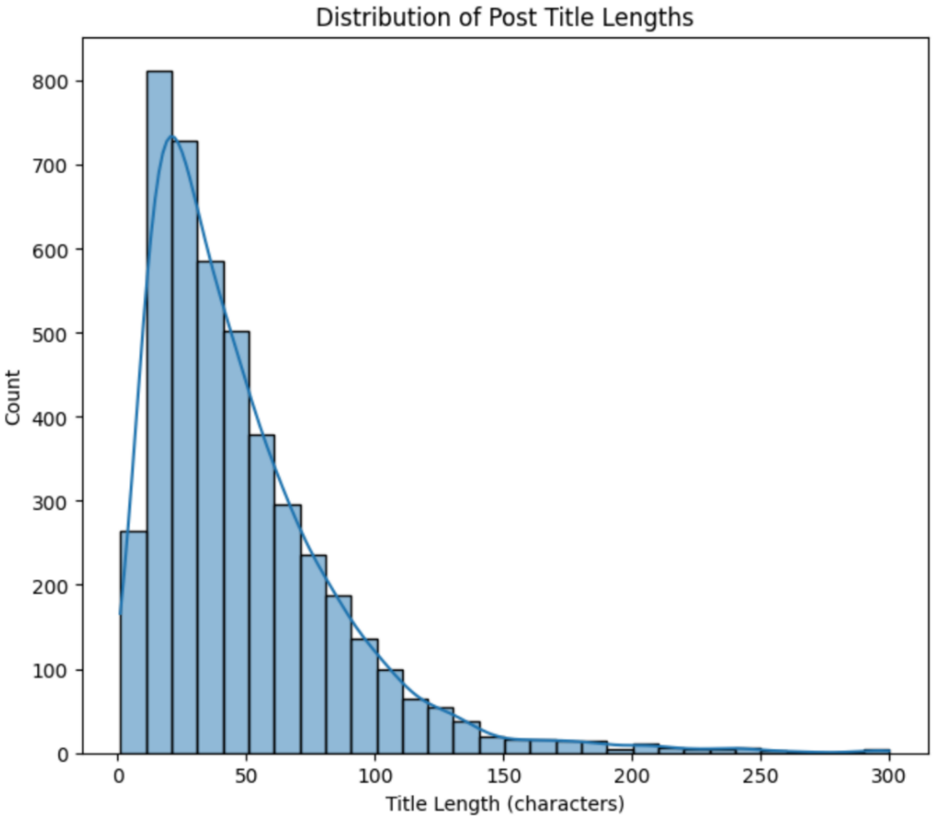

Figure 1 - Distribution of Post Title Lengths

1.3 Engagement Metrics

- Score: mean = 23,432.64, median = 452, min = 0, max = 419,309
- Number of comments: mean = 324.03, median = 14, min = 0, max = 10,637

Engagement metrics are heavily right-skewed. Most posts receive moderate attention, while a small subset achieves viral status, reflecting the typical distribution of user-generated content on Reddit.

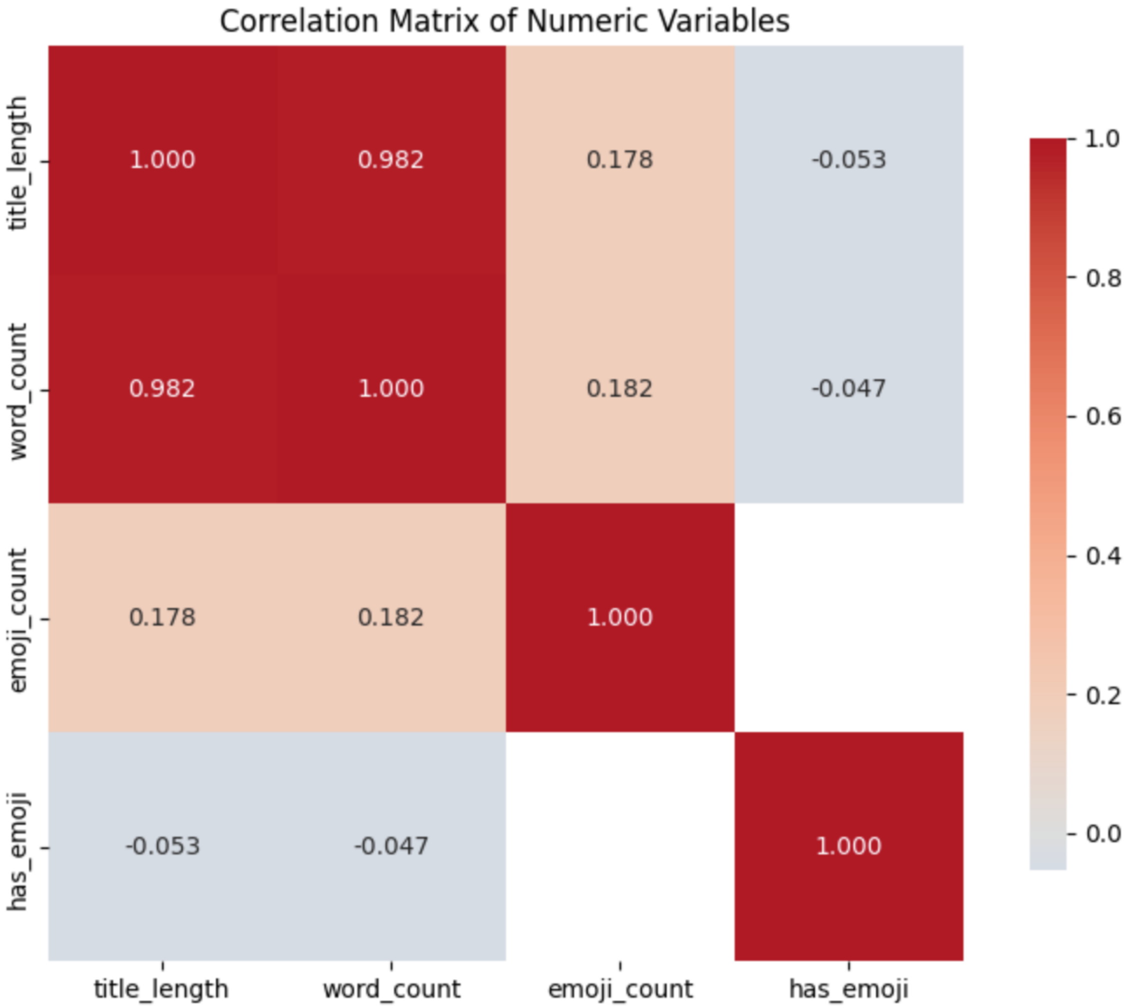

Figure 2 - Correlation Matrix of Numeric Variables

1.4 Subreddit Distribution

Top subreddits:

- r/memes: 775 posts (3.5% with emojis)
- r/funny: 760 posts (3.4% with emojis)
- r/jokes: 748 posts (0% with emojis)
- r/Oneliners: 747 posts (0.4% with emojis)
- r/standupshots: 740 posts (2.2% with emojis)
- r/cleanjokes: 738 posts (0.8% with emojis)

r/memes and r/funny show slightly higher emoji usage than other subreddits, while r/jokes has no emoji usage. Emoji use varies with subreddit norms and may reflect different stylistic conventions in humor.

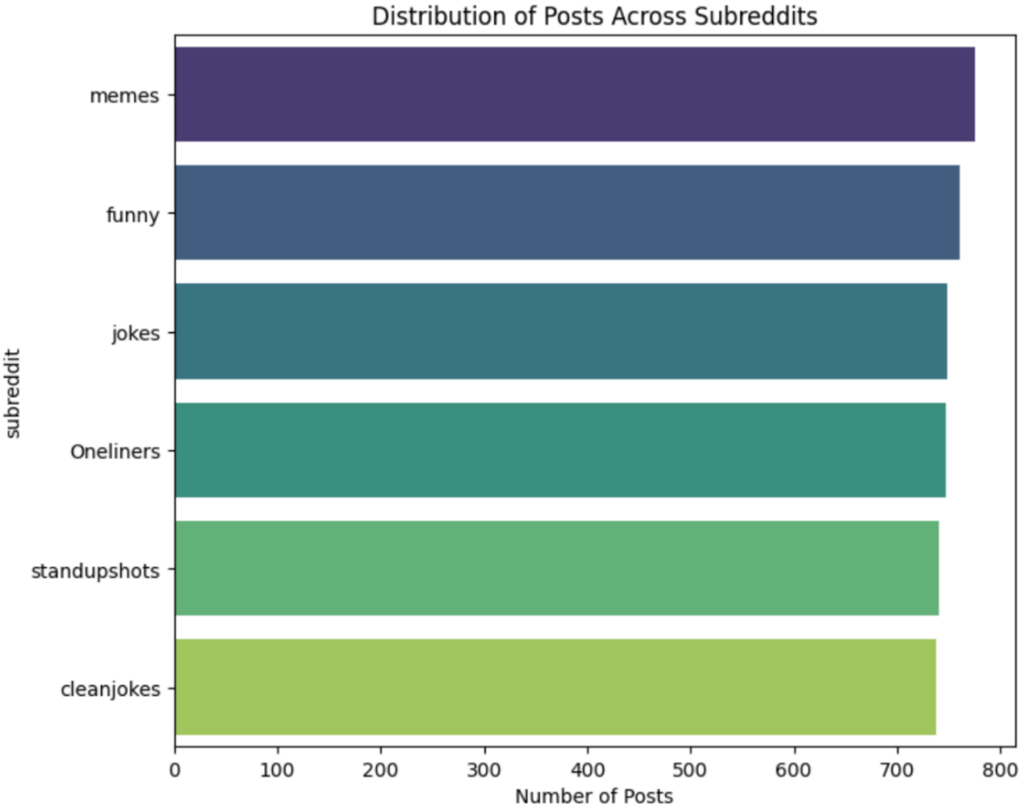

Figure 3 - Number of Posts per Subreddit

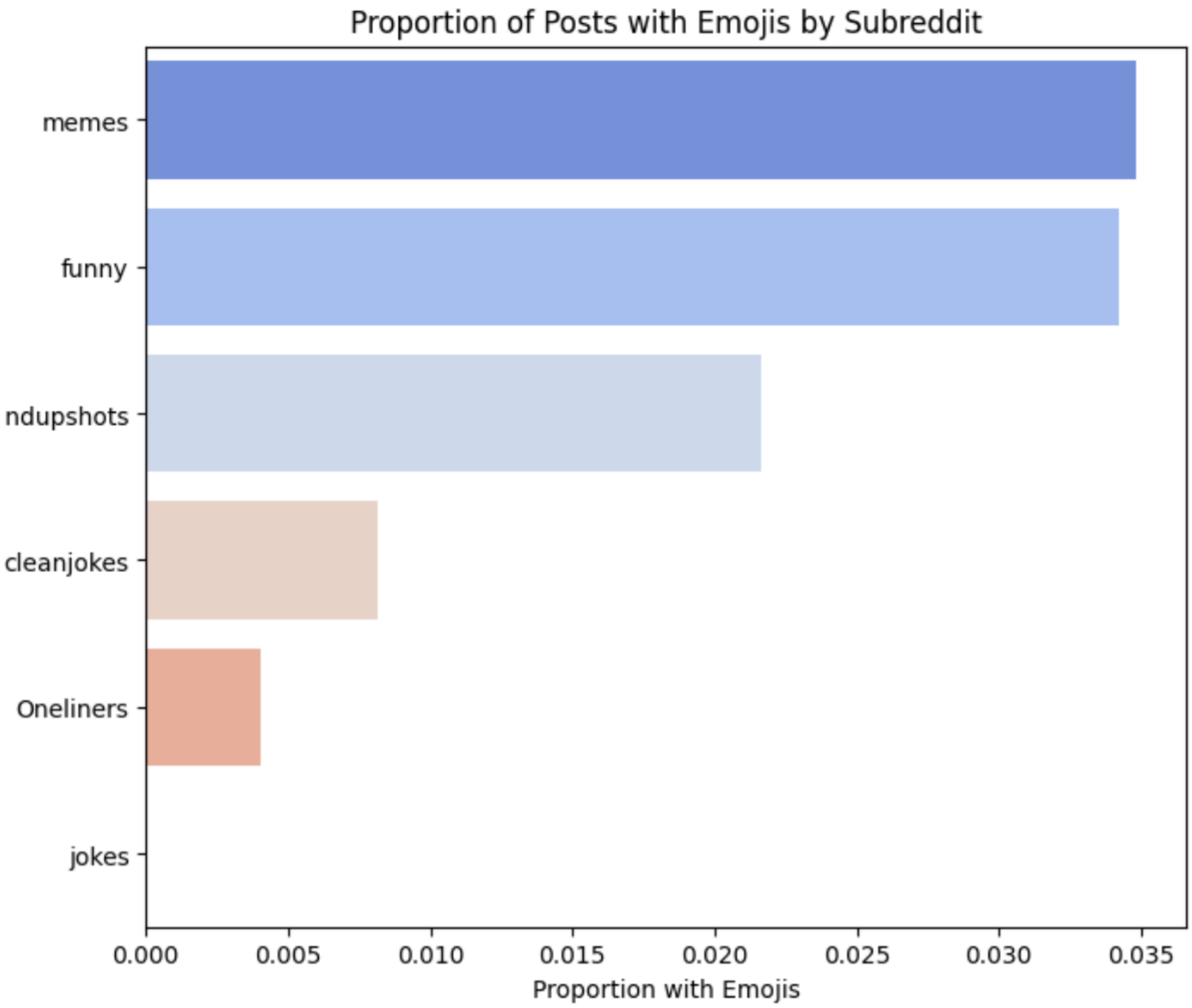

Figure 4 - Proportion of Posts with Emojis by Subreddit

1.5 Emoji Analysis

Unique emojis found: 61
Most frequent emojis:

1. 😂 (16 occurrences)
2. 😭 (16 occurrences)
3. 👀 (4 occurrences)
4. 😎 (4 occurrences)
5. ❤ (3 occurrences)

Emojis are primarily positive or expressive (😂, 😭, 👀), supporting the idea that they act as paralinguistic markers to enhance humor or express emotion. Despite their low frequency (1.7% of posts), emojis correlate with shorter headlines and potentially higher engagement.

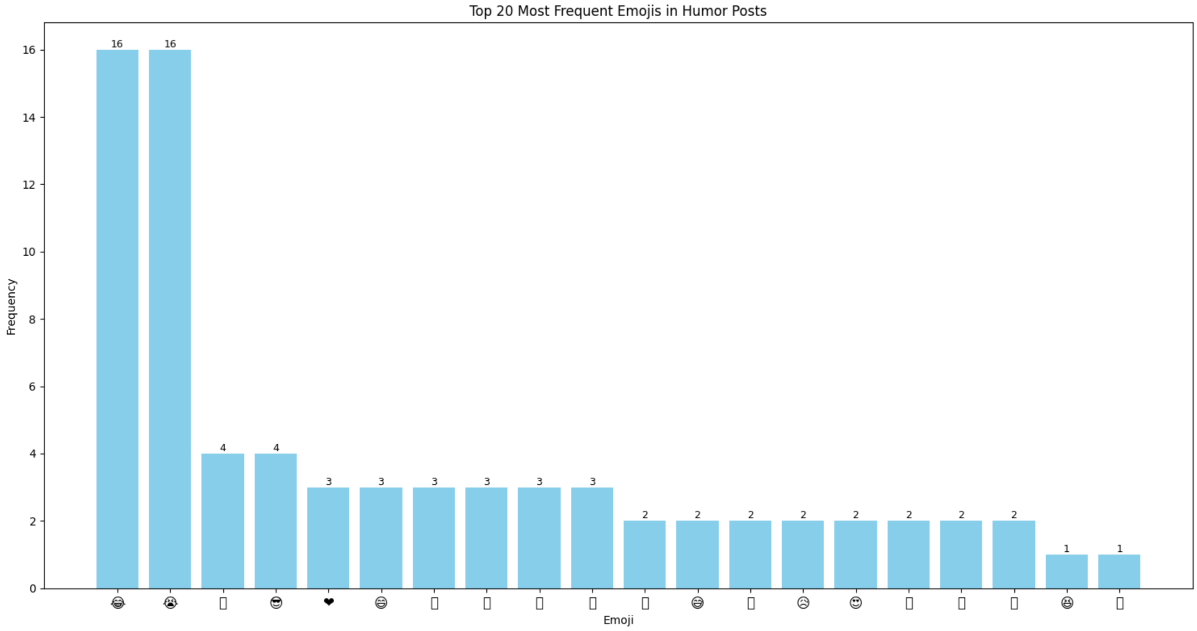

Figure 5 - Top 20 Most Frequent Emojis in Humor Posts

1.6 Key Patterns and Statistical Insights

- Posts with emojis are significantly shorter in both characters and words.
- Short, emoji-accompanied headlines suggest a style of concise, visually expressive humor.
- Engagement metrics (score, comments) are highly skewed; a small number of posts dominate attention.
- r/funny and r/memes subreddits contribute most emoji usage.

1.7 Key Insights

- Brevity as a humor cue: Short headlines dominate, with emojis often accompanying the shortest posts.
- Emojis as paralinguistic signals: Even rare, emojis like 😂 and 😭 highlight emotional emphasis in humor.
- Subreddit influence: Different subreddits show distinct emoji patterns, reflecting platform-specific stylistic norms.
- Skewed engagement: Viral posts are few but drive most interactions, emphasizing the importance of standout content in humor analysis.

2. Secondary Data

This section presents the analysis of the secondary dataset obtained from Kaggle, which contains 75,650 Reddit posts with 18 columns. The goal is to compare content characteristics, emoji usage, and engagement metrics with the primary dataset while noting any limitations inherent to pre-existing data.

2.1 Dataset Overview

- Total posts: 75,650
- Number of features: 18
- Collection period: 14-10-2025
- Duplicate rows: 0
- Missing values: High in distinguished (99.99%) and link_flair_text (83.07%), minimal in selftext (0.01%)

The dataset is large and comprehensive but contains substantial missing values for certain metadata fields, which may limit analysis of features like flair or moderation status.

2.2 Text and Content Characteristics

Table 1 - Text and Content Characteristics Table

Column	Avg. Characters	Avg. Words	Missing %
author	36.0	1.0	0
created_utc	26.0	2.0	0
distinguished	9.0	1.0	99.99
edited	26.0	2.0	0
thread_id	36.0	1.0	0
link_flair_text	6.3	1.4	83.07
selftext	206.3	38.2	0.01
title	51.8	9.9	0
extracted_utc	26.0	2.0	0

Observations:

- Headlines (title) average ~52 characters and ~10 words, similar to the primary Reddit dataset.
- Selftext fields are considerably longer, averaging 206 characters and 38 words.
- Missing values in metadata fields limit analysis of post flair or moderation indicators.

2.3 Emoji Analysis

- Posts with emojis in titles: 7,592 (~10%)
- Posts with emojis in selftext: 13,591 (~18%)
- Average emojis per post:
  - Titles: 1.43
  - Selftext: 5.26
- Total unique emojis:
  - Titles: 110
  - Selftext: 423
- Most frequent title emojis: ’, …, “, ”, ‘, é, —, £, ö,
- Most frequent selftext emojis: ’, “, ”, …, ‘, —, 🤣, –, £, 😂

Emoji usage is higher than in the primary dataset, particularly in selftext. Emojis are predominantly punctuation or expressive symbols (’ “ ” …), with a smaller fraction being traditional emojis like 😂 or 🤣.

2.4 Engagement Metrics

Table 2 - Engagement Metrics Table

Metric	Min	Median	Mean	Max	Std. Dev
num_comments	0	2	14.87	8,011	72.47
score	0	3	250.58	53,635	1,441.13
upvote_ratio	0.02	0.69	0.66	1.00	0.24

Observation:
Engagement is heavily skewed, with most posts receiving few votes or comments, while a small fraction achieves viral status. This pattern mirrors that of the primary dataset.

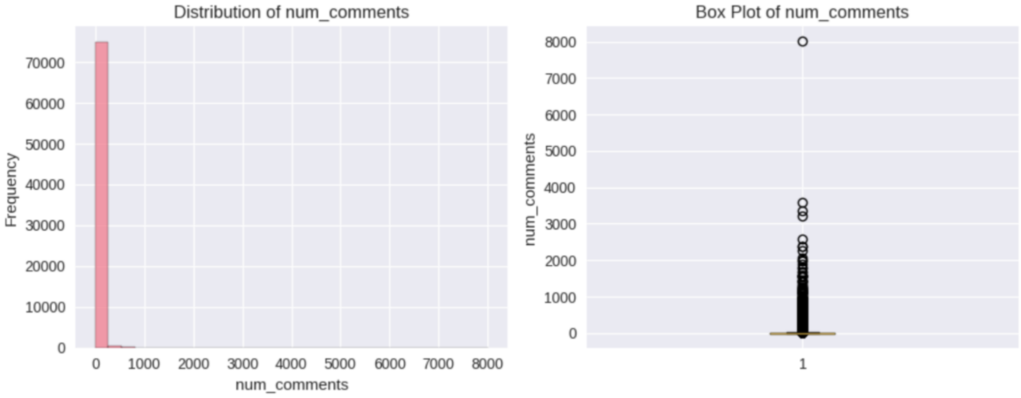

Figure 6 - num_comments Distribution and Box Plot

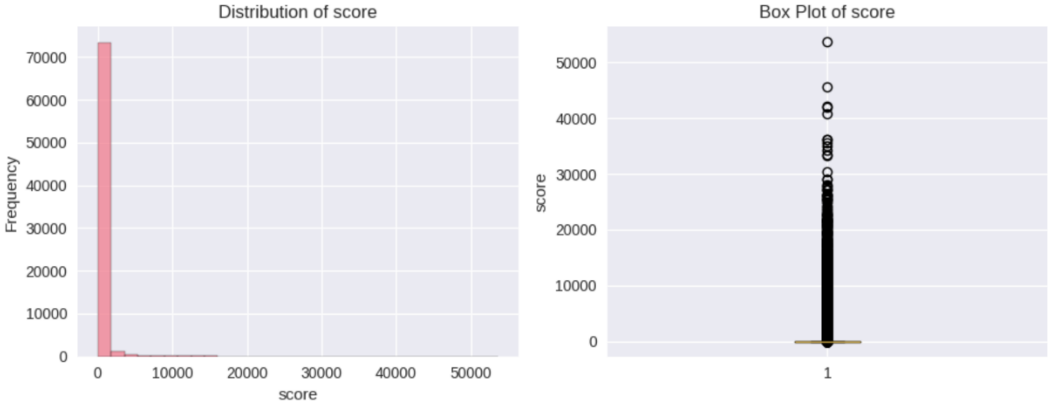

Figure 7 - Score Distribution and Box Plot

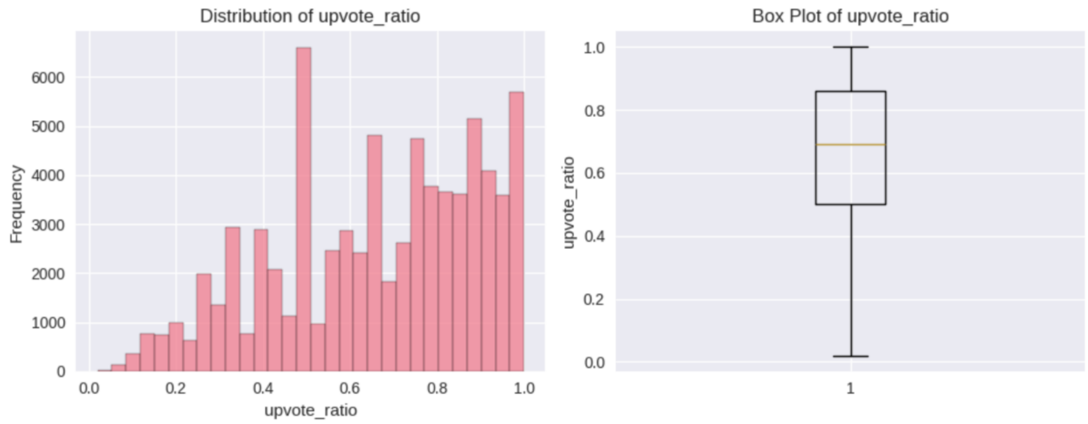

Figure 8 - upvote_ratio Distribution and Box Plot

2.5 Correlation Analysis

Strong correlations identified:

- num_comments – score: r = 0.722
- title_length – title_word_count: r = 0.972
- selftext_length – selftext_word_count: r = 0.982
- link_flair_text_length – link_flair_text_word_count: r = 0.925

Length in characters is strongly predictive of word count across all text fields. Score and comment count are positively correlated, indicating that more popular posts tend to receive more discussion.

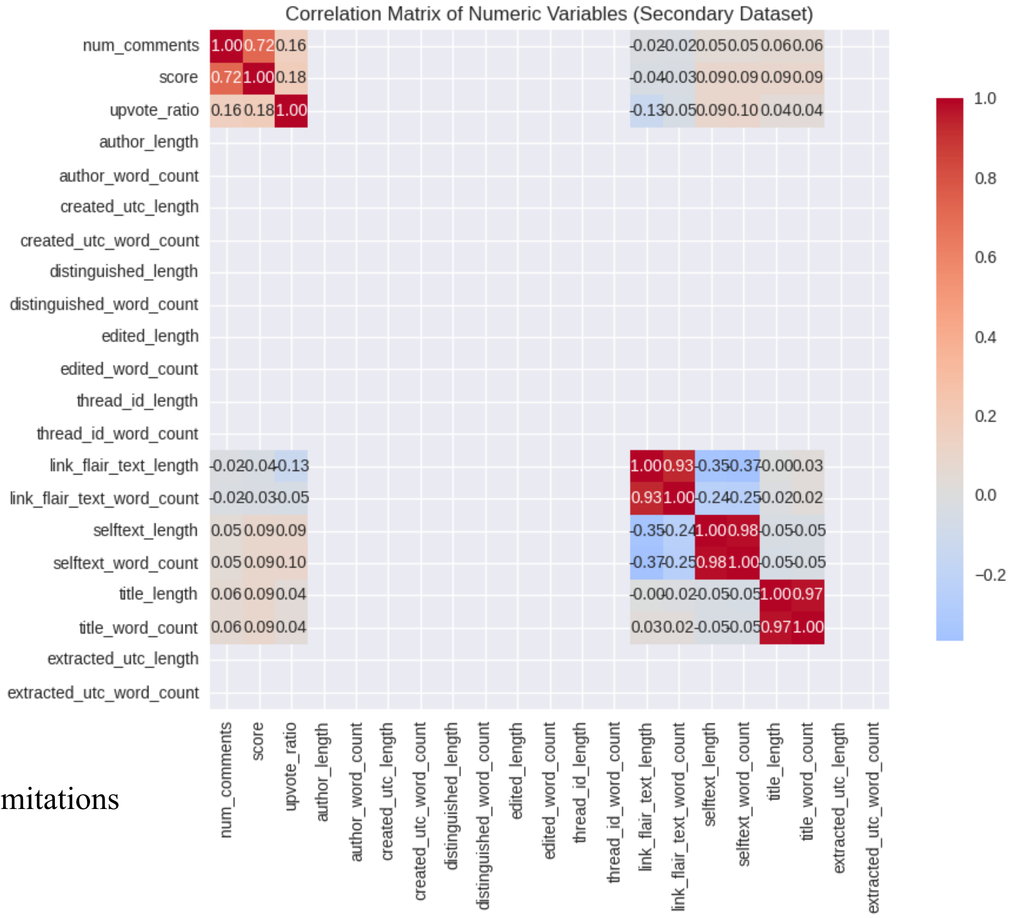

Figure 9 - Correlation Matrix of Numeric Variables of Secondary Data

2.6 Limitations

- Missing metadata: High proportions of missing distinguished and link_flair_text reduce reliability for moderation/flair-based analysis.
- Sampling bias: Dataset may over-represent certain joke types or curated examples from Reddit.
- Temporal information unknown: Data collection period is not fully specified, making comparisons with primary data potentially time-inconsistent.
- Emoji distribution: Some "emojis" are punctuation marks, so analysis of actual expressive emojis requires careful filtering.

2.7 Key Insights

- Headlines are similar in length and word count to the primary dataset.
- Emoji usage is more frequent in selftext than in titles (~18% vs 10%).
- Engagement is highly skewed, consistent with viral content dynamics.
- Strong correlations exist between post length and word count, and between score and comment count.
- Limitations of metadata and sampling bias should be considered in subsequent analysis.

3. Comparison Between Primary and Secondary Datasets

This section compares the primary Reddit dataset with the secondary Kaggle dataset, highlighting similarities, differences, and contextual insights.

3.1 Key Metrics and Text Characteristics

Table 3 - Key Metrics Table

Metric	Primary Dataset	Secondary Dataset	Interpretation
Dataset Size	4,508 posts	75,650 posts	Secondary dataset provides more statistical power
Content Type	Diverse humor (memes, oneliners, standup)	Traditional jokes	Primary covers broader humor, secondary is focused
Emoji Usage in Titles	1.7%	10.0%	Community norms differ; r/Jokes uses more emojis
Average Title Length	48.2 characters	51.8 characters	Titles are slightly longer in secondary dataset
Average Words per Title	9.2 words	9.9 words	Similar complexity in joke setup
Emoji Diversity	61 unique emojis	110 unique emojis	Secondary dataset shows wider emoji variety

Both datasets show similar title lengths and word counts, but emoji usage and dataset scale differ significantly.

3.2 Patterns, Discrepancies, and Biases

Consistent Patterns:

- Humor content is present in all communities studied.
- Title lengths and word counts are comparable.
- Engagement metrics show skewed distributions (few highly upvoted posts).

Divergent Patterns:

- Emoji usage: 1.7% (primary) vs 10% (secondary)
- Dataset size: 4.5K vs 75K posts
- Content focus: Broad humor (primary) vs traditional jokes (secondary)

Possible Explanations:

- Community norms (r/Jokes encourages more emoji use)
- Content type differences (memes/oneliners vs structured jokes)
- Temporal differences and user demographics

Sampling Bias Considerations:

- Primary dataset: Smaller, curated, limited to six subreddits
- Secondary dataset: Larger, focused on one subreddit, less diverse in humor types

3.3 Implications for Humor Research

- Humor expression varies across Reddit communities.
- Emoji usage is context-dependent and influenced by community norms.
- Both datasets are complementary: primary provides breadth, secondary provides depth.
- Comparative analysis should consider dataset scale, content type, and subreddit focus.
- Similar text characteristics (length and word count) suggest some universal features in humor posts, while emoji usage highlights community-specific behaviors.

4. Summary of New Insights and Hypotheses

EDA revealed that humor posts across Reddit tend to have short, concise titles (around 9–10 words, 48–52 characters), suggesting brevity is a common cue for humor. Emoji usage varies by community—rare in the primary dataset (1.7%) but more common in the secondary dataset (10%)—and posts with emojis are generally shorter, indicating a link between emojis and punchy humor. Engagement is skewed, with few posts achieving high scores or comments. Based on these observations, new hypotheses include: emoji presence may increase engagement, shorter titles are more shareable, and certain lexical patterns and punctuation may predict humor. These insights guide further testing and modeling in Phase 3.

Phase 3

1. Overview

The objective of Phase 3 was to develop predictive models and sociolinguistic insights for humor detection using a large corpus of Reddit humor content and an auxiliary short-jokes dataset. This phase focused on:

1. Humor Classification: Predicting whether a Reddit post is humorous (Label = 1) or non-humorous (Label = 0).
2. Sociolinguistic Analysis: Understanding how humor varies by community, emoji usage, text length, and linguistic style.

Multiple models were developed and evaluated (Logistic Regression and an RNN neural model) using a combination of engineered features and deep text representations. All modelling decisions, evaluations, and interpretations are documented below.

Note on Data Usage in Phase 3:
In Phase 3, the secondary dataset used for modeling and sociolinguistic analysis differs from the Kaggle dataset referenced in Phase 2 for comparative analysis. For this phase, we incorporated the Hugging Face Short Jokes Dataset (containing ~231,657 short jokes) to enrich the training data and enhance the vocabulary for neural models. This allowed for a more robust exploration of humor classification and linguistic patterns, complementing the primary Reddit dataset used throughout the project.

2. Data Preparation for Modelling

2.1 Final Modelling Dataset

After cleaning, filtering malformed entries, and incorporating linguistic features, the dataset was significantly expanded in size to improve model training and capture a broader range of humor patterns.

- Final dataset size: 20,804 Reddit humor-related posts
- Columns included:
  ['subreddit', 'type', 'title', 'score', 'num_comments', 'emojis', 'has_emoji', ... engineered features]
- Label distribution:
  - Humor (1): 14,579 posts
  - Non-humor (0): 6,225 posts

The dataset also integrated 231,657 short jokes to enhance vocabulary learning for neural models.

Total dataset for modelling: 20,804 entries with 10 final features.

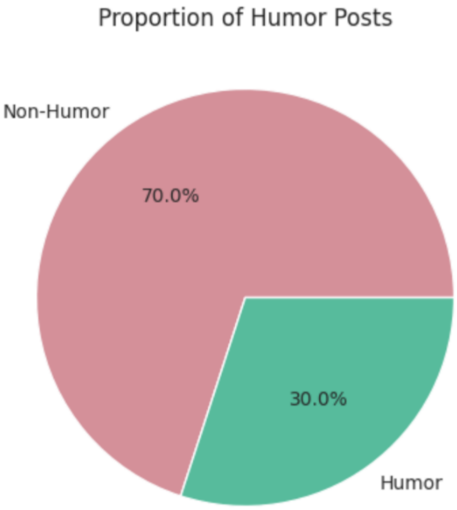

Figure 10 - Proportion of Humor Posts

3. Task 1: Humor Classification

3.1 Baseline Model

A Dummy Classifier predicting the majority class (humor):

- Accuracy: 0.70
- Precision: 0.51
- Recall: 1.00
- F1-Score: 0.68

As expected, this baseline offered no meaningful predictive capability beyond skewed prior distribution.

3.2 Logistic Regression Model

A linear model using TF-IDF features and engineered sociolinguistic variables.
Performance (Test Set):

- Accuracy: 0.6353
- Precision (Humor): 0.76
- Recall (Humor): 0.70
- F1-Score: 0.6439

Class-level performance:

Table 4 - Logistic Regression Class-level Performance

Label	Precision	Recall	F1	Support
0	0.41	0.49	0.45	1,862
1	0.76	0.70	0.73	4,381

Interpretation:
Logistic Regression captures some humor patterns but struggles with non-humor detection due to class imbalance and nonlinear language patterns.

3.3 RNN Model

To capture sequential linguistic cues, an RNN was trained with:

- Enhanced vocabulary from 231k short jokes
- Class weighting to address imbalance

Performance:

- Accuracy: 0.6277
- F1-Score: 0.6396

Class-level performance:

Table 5 - RNN Class-level Performance

Label	Precision	Recall	F1	Support
0	0.41	0.53	0.46	1,862
1	0.77	0.67	0.72	4,381

Training dynamics indicated overfitting after epoch 3, with validation performance plateauing.

Interpretation:
The RNN better captures creative or unusual humor structures but struggles on shorter or ambiguous texts.

3.4 Model Selection for Humor Classification

Although the RNN is more expressive, Logistic Regression slightly outperformed it in overall accuracy and stability.

Final Choice: Logistic Regression

Reasons:

- More stable across validation folds
- Less prone to overfitting
- Comparable or superior F1-score
- Faster, easier to interpret

4. Sociolinguistic Analysis

This section replaces the earlier engagement-prediction task since the updated dataset contains engagement metrics but labels are not reformulated for a prediction task. Instead, a deeper community-level humor and style analysis was performed.

4.1 Subreddit Humor Patterns

A community-level breakdown revealed strong differences in humor density, emoji usage, and engagement:

- Subreddits like unexpected, TIHI, showerthoughts, and ContagiousLaughter exhibit near-universal humor labeling (0.98–0.99).
- Communities like pun (0.07) and lamejokes (0.01) show extremely low detected humor, suggesting semantic ambiguity or model mismatch.

Engagement indicators (average score and comments) varied widely, with funny, unexpected, and ProgrammerHumor leading in interactions.

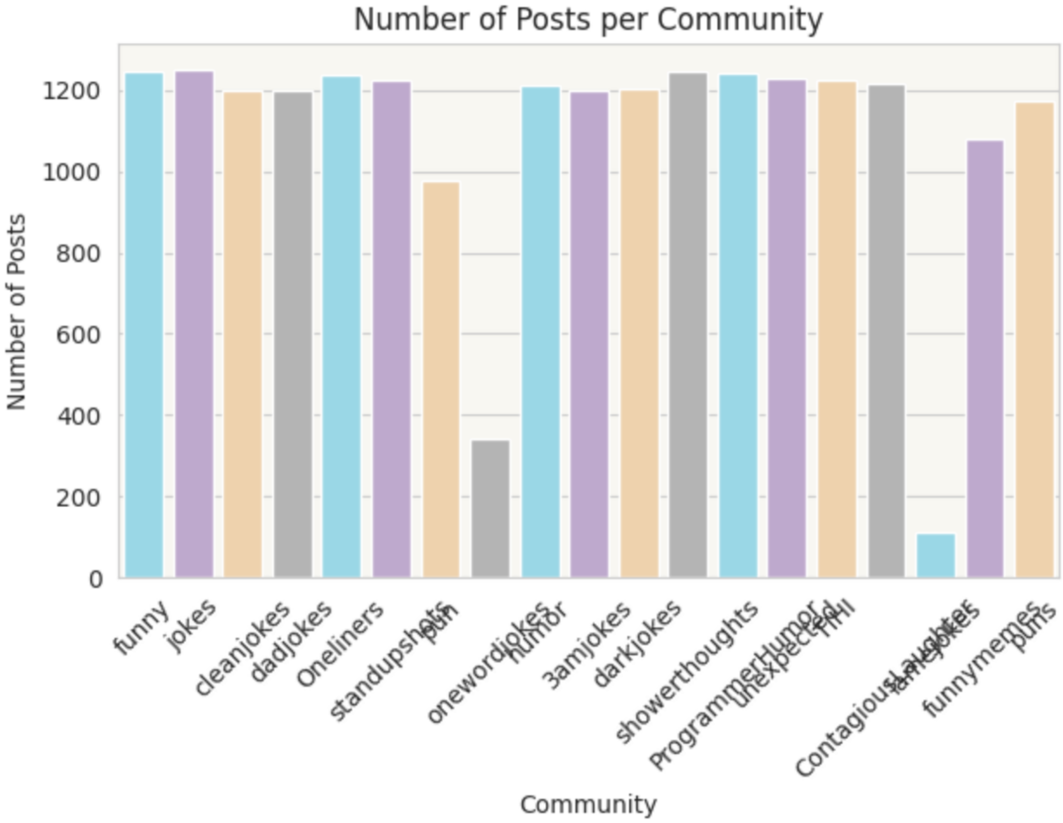

Figure 11 - Number of Posts Per Community

4.2 Emoji Impact on Humor

Emoji presence was rare (only 475 posts), but:

Table 6 - Emoji Impact on Humor

Emoji?	Humor Rate	Avg Score	Avg Comments
No	0.701	10,081	182.7
Yes	0.709	7,581	136.9

Result:
Emoji usage has minimal impact on humor labeling and correlates slightly negatively with engagement.

4.3 Text Length Analysis

Character length varied dramatically by subreddit:

- Shortest: onewordjokes (7.3 chars avg)
- Longest: showerthoughts (114.5 chars avg)

Comparison to short-jokes dataset:

Table 7 - Short Jokes vs Reddit Humor Datasets

Dataset	Avg Length	Standard Deviation
Short jokes	93 chars	35.3
Reddit humor	46.1 chars	36.8

Interpretation:
Reddit humor is typically shorter, punchier, and more informal than traditional joke structures.

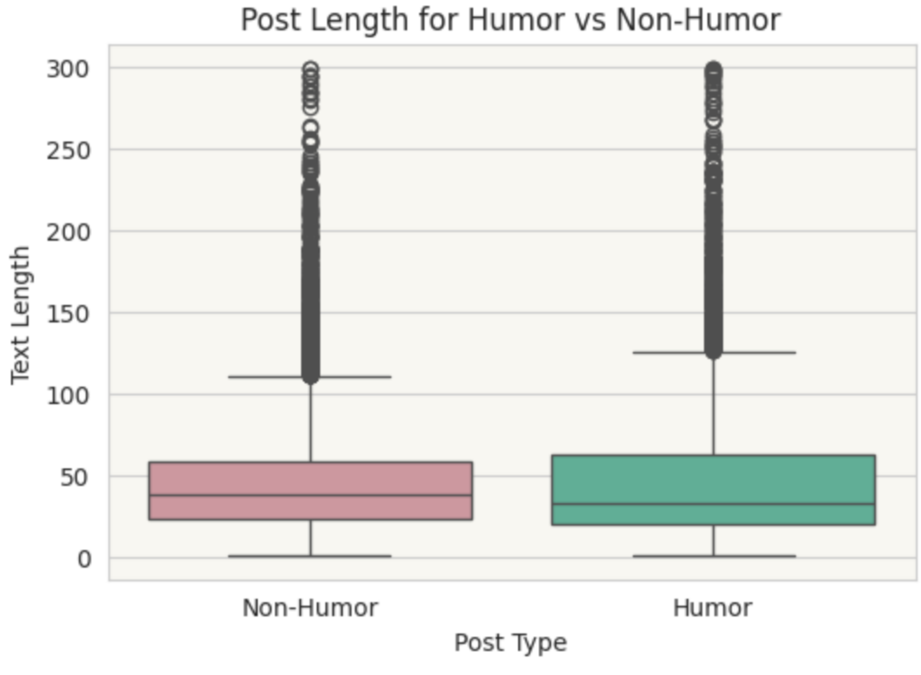

Figure 12 - Post Length Humor vs Non-Humor

4.4 Linguistic Insights

Patterns emerging from feature analysis:

- Humor correlates with brevity, informal wording, and community norms.
- Non-humor often includes ambiguous short posts (“Fire at will”) or factual statements.
- Informal vernacular (AAVE, slang like *no cap*, *bussin*) is consistently labelled humorous.

5. Communication of Results

Key Findings

- Humor labeling is highly dependent on community context.
- Emoji usage barely influences humor classification.
- Reddit humor is significantly shorter than classic joke structures.
- Logistic Regression performs slightly better than the RNN despite the neural model's theoretical advantages.
- Both models show difficulty detecting non-humor due to dataset imbalance and the wide creativity of humor expression.

6. Model Summary

Table 8 - Model Summary

Task	Best Model	Accuracy	Precision	Recall	F1
Humor Classification	Logistic Regression	0.635	0.66 (weighted)	0.64 (weighted)	0.64
Secondary Model	RNN	0.627	0.66 (weighted)	0.63 (weighted)	0.6396

Conclusion

This project aimed to answer the question: “How do language, context, and social factors influence the perception and structure of humor in short textual content, such as headlines?” Across all three phases, the findings show that humor online is shaped by both linguistic features and the social environments in which it appears.

Phase 1 created a diverse dataset by combining curated short jokes with real Reddit posts. This revealed early differences in how humor is expressed across communities and showed that cultural and platform biases shape what kinds of humor appear.

Phase 2 explored the text in detail. Humor headlines were generally short, informal, and occasionally supported by emojis. Different subreddits favored different humor styles, which shows that audience expectations and community norms strongly influence how humor is written and interpreted. Engagement patterns also showed that a small number of posts receive most of the attention, which highlights the social nature of humor reception.

Phase 3 applied machine learning models to classify humor. Logistic Regression and an RNN were able to capture some humor cues, although both struggled with ambiguity and class imbalance. The results supported earlier insights by showing that features such as brevity, informal language, and community-specific styles help predict humor, while deeper contextual meaning remains difficult for models to learn.

Overall, the project shows that language, context, and social factors work together to shape humor. Linguistic choices such as short length, wordplay, and informal vocabulary contribute to humorous expression. Context, including subreddit culture and shared references, determines how audiences interpret a joke. Social factors, such as community norms and interaction patterns, influence which humorous posts gain attention.

In summary, humor in short online headlines is not only a matter of what is written. It also depends on where it is posted, who reads it, and the shared cultural knowledge that gives a headline its humorous meaning.