CS-401

Applied Data Analysis

P1 Milestone

Authors

Amélie Menoud
András Horkay
Cyrielle Manissadjian
Dominic Stratila
Katia Todorov

Andras's Ideas

I am going to put my submission here so that when we compare and pick the project we will do, we can compare them easily

When I Grow Up, I'll Be Anything but a Politician: Professions and Social Attitudes in the New Yorker Caption Contest

This project explores how different professions are portrayed in the New Yorker Caption Contest and how audiences respond to them. While some jobs, such as doctors, are often admired, others, like politicians and lawyers, are frequent targets of humour. Our research will focus on, but might not be limited to: Which professions appear most often in captions? Are certain occupations consistently rated funnier? Are some portrayed in a positive or negative light? To address these questions, we will first identify professions mentioned in captions using a U.S. Census occupation list, accounting also for indirect workplace references (e.g., "courtroom" for lawyers). The professions will then be grouped into broader categories (healthcare, law, education, politics, etc.). Finally, we will analyse their frequency, average funniness scores, and recurring stereotypes. We hope the findings will shed light on societal attitudes towards various professions as reflected in popular culture, and perhaps give ideas which professions to avoid in future careers to avoid being the butt of jokes.

Laughing at the Forbidden: Taboo Themes in the New Yorker Caption Contest

This project investigates how taboo themes appear in the New Yorker Caption Contest and how audiences respond to them. While taboo jokes may be funny due to shock value, they can also be offensive or harmful. Some key questions we will answer are: Which taboo topics occur most often? How are they rated in terms of funniness? Do their frequencies change over time or with the political landscape in the US? To answer these, we will define taboo categories (e.g., suicide/self-harm, racism, sexism, violence, religion) and identify their presence in captions through keywords and phrases. We will then analyse the frequency at which these topics occur, their associated funniness scores, and any temporal trends, paying attention to events such as elections or economic crises that may influence prevalence. The project will shed light on how societal attitudes towards taboo topics are reflected in humour and how audience perceptions of these themes evolve over time.

Funny or Stereotypical? Gender Roles in the New Yorker Caption Contest

This project examines how gender is represented in the New Yorker Caption Contest, both in cartoons and in audience-submitted captions. The key questions we will attempt to answer are: Do men appear more frequently than women in the images? When women are depicted, are they more often shown in stereotypical roles (domestic or caregiving) rather than professional settings (as leaders)? In captions, are gendered terms or phrases used that reinforce stereotypes? We will analyse cartoon metadata to measure representation, classify roles to detect stereotypes, and examine captions for gendered language. These steps may involve image as well as text analysis. Finally, we will compare funniness scores to assess whether captions using gendered language or stereotypes are rated differently by audiences. The results will provide insight into the persistence of gender stereotypes in humour and their reception by the public.

Cyrielle's Ideas

Why cognitive biases make jokes funnier ?

Cognitive biases are systematic deviations in thinking distorting our perception, judgment, decision-making. Around 250 cognitive biases are referenced, they are classified into 6 categories, we study 3 of them.

- Attentional biases: Do raters become less amused after seeing too many captions ("bored-rater effect") ? We can track the vote distributions over ranking order to detect potential decrease in average scores. Do certain visual elements (such as anthropogenic animals, uncanny objects/situations) attract more humorous interpretations ? Using the image metadata (objects, locations, entities) to test whether attention-grabbing visuals correlate with higher fun.

- Judgment biases: Is humor works with left-brain/right-brain ? In other word, do logical/emotional joke produces more/less fun ? Use a text-based metric to measure polarity and subjectivity of the caption might help to answer. Calculate statistics per identified groups.

- Cultural biases: "Do you get it ?" is an often-heard sentence when talking about humorous content. Do culturally referential captions (brands, celebrities, memes) outperform classical ones ? Are winning captions influenced by events ? We can detect proper nouns, cultural entities in captions with text processing, then compare funniness scores via T-tests. Overlay with publication dates (an other dataset need to be merged) to match historical-political context.

Are we all sheep and do we reproduce the same patterns?

- Diversity across cartoons : Do we keep writing the same jokes forever ? Do trends come back (as in the world of fashion and/or music) ? Do certain humor types reappear over time regardless of image content ? We can extract dominant joke types across years (e.g., existential crisis, sarcasm, animals acting human) using thematic clustering, then correlate with cartoon metadata.

- Diversity within cartoons : Are thousands of captions mainly identical ? Do most people have the same interpretation of the cartoon ? Calculating intra-cartoon diversity using semantic distance metrics, and label clusters with classic humor theory types to detect which "family of humor" dominates and whether originality helps to win the contest.

Do all caption submissions really have a possibility to win ?

- About fun : Is there a formula for "winning" caption ? Does good writing comes from good ideas or rather good sentences ? Are funny captions sharp, absurd, readable, abstract . . . ? For each cartoon, we can group captions by humor strategy (statement vs dialogue, clever vs literal), compare pairwise caption clusters telling the "same idea" but with different phrasing. We can also compare cross-cartoon clustering, by plotting a relation between the captions similar to odd elements (uncanny description) or similar to the expected context (description).

- About the System : Is the game fair or algorithmically set up ? Perhaps some captions are highlighted by the competition (reputation bias: Proposals that receive feedback early on accumulate more). We can plot funniness scores vs number of votes to detect power-law effects. We also can play with comparison of low-vote/high-score vs high-vote/moderate-score captions.

Dominic's Ideas

Causality or Correlation? Unlikely Trends in the New Yorker Caption Contest

This project explores strange or unexpected correlations between different trends in the Caption Contest that appear unrelated at first sight. For example, Is there a connection between the seasons and jokes about cats? How do jokes about politicians affect the number of people depicted in the cartoon? Does the presence of certain objects (like chairs, doors, or telephones) correlate with particular humor strategies? We will systematically identify such patterns by cross-analysing captions, metadata, and funniness scores across distinct spheres (visual elements, themes, cultural references). The focus is not necessarily to establish causal relationships, but rather to highlight these surprising overlaps and ask whether they reflect deeper societal attitudes, hidden cognitive associations, or simply coincidences. The outcome may provide both serious insights into how people connect ideas in humor and amusing examples of "funny but meaningless" correlations.

Katia's Ideas

Humor in context: How news shapes our fun

How does the perception of humor in captions change over time, and how is it influenced by current events?

This project aims to identify periods in which similar types of humor are produced independently of the image content and to determine when they are perceived as funnier. These patterns will then be analyzed in relation to the global context, such as political, environmental, or societal events.

To answer this questions, the following steps could be followed : Gather information about the dates for each image; Implement text analysis on caption; Create clusters of common humor types; Identify periods and trends showing particular behavior; Link the findings to current events, including health data, political context, and global events.

Creativity is the key

How does the creativity of a caption influence its popularity and is creativity inhanced by certain themes?

This project focuses on understanding how creative captions behave, analyzing both the total number of captions and how closely the words relate to the content of the image. It will also looks at why certains images will attract more captions than other. The distinction between obvious and subtle humor could be explored if times allow it.

Steps: Perform text analysis: extract the most frequently used words, removing stopwords; Analyze captions within images: examine whether captions that use words similar to the image description tend to receive higher scores; Analyze captions between images: study differences in themes, identify which images attract more captions, which are more popular for voting, and which are perceived as funniest; Interpret the results and draw conclusions.

What fun is about

This project aims to capture, if it exists, the underlying patterns of the funniest captions. It will explore different aspects of semantics, including the lexicon, sentence structure, interpretation of quantifiers, and the use of positive or negative phrasing.

Steps: Feature extraction (analyze captions for textual features such as length, word choice, sentence structure, use of quantifiers, and positive/negative phrasing, etc.); Create a dataset with these features as columns for each caption; Apply multivariate techniques (e.g., PCA) to identify which features most strongly correlate with perceived funniness; Investigate the key variables identified to understand their influence on fun perception; Draw conclusions. If time allows: explore patterns in votes versus fun ratings and investigate why they behave like this (fun: exponential, neutral: linear, not funny: log) and whether this behavior observed in some images is the norm or a special case.

Amelie's Ideas

1. Tops and Flops: Mapping humor styles Here the goal is to build a typology of humor in the New Yorker Caption Contest. We would try to describe the diversity of humor styles and their prevalence. Do certain categories such as wordplay, dark humor, irony, absurdism... dominate? Can we cluster captions into clear stylistic groups, and how stable are these clusters across time? Once we have that, we could see if some humor styles tend to appear only occasionally, while others are overrepresented. A further step would be to analyze wether certain images encourage specific types of humor, for instance wether office settings atttract political jokes or surreal drawings trigger absurd responses. I think that combining clustering methods, linguistic feature extraction and doing the analysis with cartoon types will produce a kind of ”map of humor”.

2. Can we laugh at everything? A cultural lens of humor This topic would approach the dataset as a reflection of cultural and social dynamics. The central question is not just which captions win, but what they reveal about the humor in the american society. Do taboo or dark jokes ever make it far, or are they always rejected? Do captions mirror political or cultural event such as the elections, economic crisis, wars... ? An additional angle is sentiment: have captions grown more cynical, angry or ironic over time, or do they tend to remain playful and light? Another intersting point would be universality. Styles like absurd humor may travel weel, while dense puns or niche cultural references might struggle because they are not accessible to all. The project would compare how the crowd and the editors handle sensitive humor. Through sentiment analysis, topic modeling and contextualisation with real-world event, this project would highlight humor as a cultural barometer rather than just a contest.

3. How to win the caption contest: A predictive recipe Here, we want to test wether it is possible to predict contest winners and finalists: Can we quantify humor well enough to model editorial decisions ? Features such as caption length, readability, syntax and the presence of a punchline could be extracted to train a classifier. We could also measure staying close to the cartoon incresases chances of success. Another angle is dynamics: do early votes snowball into popularity, and does voter fatigue bias later judgments? Finally some participants win multiple times, do they have a distinctive style that we can detect ? By building and evaluating predictive models while also interpreting the key features, this project would produce a playful yet data-driven “guide to winning the New Yorker Caption Contest.”